Aberrantly Expressed CeRNAs Account for Missing Genomic Variability of Cancer Genes via MicroRNA-Mediated Interactions

Hua-Sheng Chiu

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2014

ABSTRACT

Aberrantly Expressed CeRNAs Account for Missing Genomic Variability of Cancer Genes via MicroRNA-Mediated Interactions

Hua-Sheng Chiu

There is growing evidence that RNAs compete for binding and regulation by a finite pool of microRNAs (miRs), thus regulating each other through a competing endogenous RNA (ceRNA) mechanism. My dissertation work focused on systematically studying ceRNA interactions in cancer by reverse-engineering context-specific miR-RNA interactions and ceRNA regulatory interactions across multiple tumor types and study the effects of these interactions in cancer. I attempted to use ceRNA interactions to explain how genetic and epigenetic alterations are propagated to target established drivers of tumorigenesis. Using bioinformatics analysis of primary tumor samples and experimental validation in cell lines, I have investigated the roles that mRNAs and noncoding RNAs can play in tumorigenesis via ceRNA interactions. Specifically, I studied how RNAs target tumor-suppressors and oncogenes as ceRNAs, and attempted to accounting for some of the missing genomic variability in tumors.

Table of Contents

List of Figures and Tables

Acknowledgements

I would like to thank my wonderful parents, Po-Hung Chiu and Chi-Mei Weng, brother, Hua-Yueh Chiu, and my wonderful wife, Hui-Fen Kao for supporting me through my Ph.D. studies. Without your encouragement and totally selfless giving, I will not be able to smoothly graduate in four years.

Personally I would like to express my sincerest appreciation and thank to Dr. Pavel Sumazin and Dr. Mukesh Bansal for giving me insightful advice and guidance on how to strengthen my research skills. I truly appreciate your time and energy throughout my Ph.D. studies at Columbia University. Thank you very much for treating me with respect and being a friend throughout my learning experience.

Thank you to Dr. Andrea Califano, my thesis advisor, for creating a wonderful research environment and teaching me what you have known and seen about academia. Your valuable suggestions are laying a solid foundation for my future job search. I simply could not ask for a better thesis advisor.

Finally, my appreciation goes to Dr. Eric Schadt, Dr. Gustavo A. Stolovitzky, Dr. Raul Rabadan, and Dr. Pavel Sumazin for agreeing to serve on my thesis committees. Your valuable comments bring me better problems and a better thesis story.

Dedication

To my parents and wife

Chapter 1: Introduction

1.1 Preface

MicroRNAs (miRNAs or miRs) are small non-coding RNAs of about 22 nucleotides that bind to partially complementary sites in their target RNAs, directly inducing RNA degradation and mRNA translational repression [1-2]. A growing body of evidence has linked miRs to tumorigenesis and tumor progression, suggesting their potential value as biomarkers and as targets for therapeutic intervention. Currently, miRs are known to regulate tumor cell growth [3-6], and their expression profiles are used to classify tumors [7-8] and to differentiate between molecular tumor subtypes [9]. Conversely, mRNAs have long been thought to be passive carriers of genetic information and their regulatory roles as RNAs in normal biological process and development of disease such as cancer were largely disregarded by scientists.

Working in parallel, the Pandolfi and Califano labs have recently uncovered a new post-transcriptional regulation layer called the competing endogenous RNA (ceRNA) regulatory network (or ceRNET) [10-13] or the miR program-mediated regulatory (mPR) network [14]. Studying regulation in glioblastoma, I have shown that mRNAs can regulate, and be regulated by, other mRNAs by competing for their shared miRNA regulators [14]. The rationale behind the ceRNET is that when two mRNAs share a common set of miR regulators, increases in the number of transcripts of one mRNA will recruit (or sponge up) more of the available miRs and induce corresponding increases in the number of translatable transcripts of the other mRNA, and *vice versa* [15]; see Figure 1.1. On a genome-wide basis, I have shown that these interactions establish a novel, large-scale regulatory layer and raise the question of the potential role of ceRNAs in tumorigenesis [14].

The ability of ceRNAs to regulate RNA turnover by titrating their shared miRNAs was first reported in plants [16] and later in human tumors [10,17]. Recently, ceRNAs have since been implicated in the pathogenic dysregulation of *FOXO1* in HVS-transformed T cells [18], of *MAML1* and *MEF2C* in muscle tissues [19], and of *PTEN* in prostate cancer [10], glioma [14] and melanoma [13]. While these reports suggest that ceRNA regulation affects development and disease, whether ceRNAs may constitute a relevant mechanism in cellular pathophysiology, and especially in tumor etiology, remains the subject of significant controversy [20].

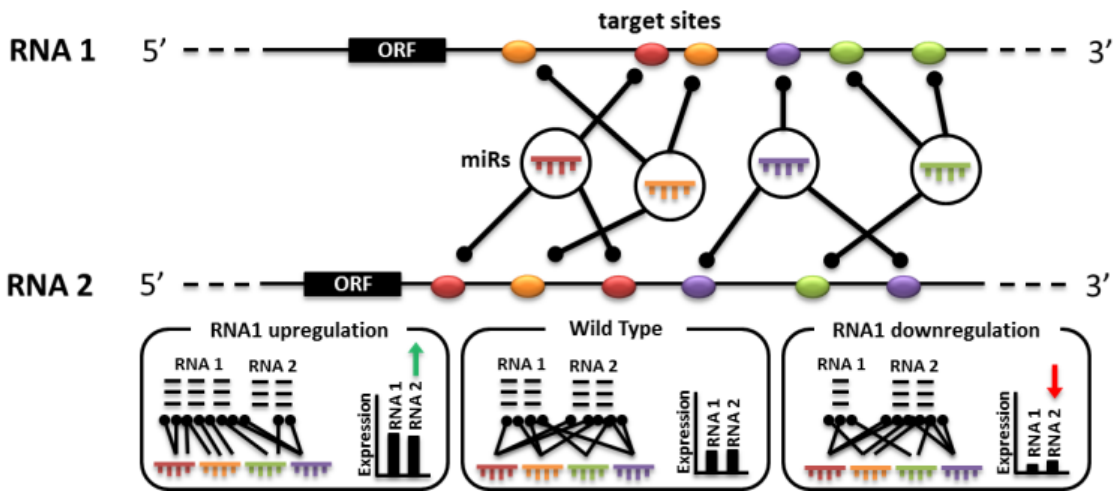


Figure 1.1. Model for ceRNA regulation

Genes whose transcripts compete for common miRNA regulators up and down regulate one another through miRNA titration. Up regulation of RNA 1 sequesters common miRNA regulators, leading to weaker down regulation of RNA 2 transcripts. Analogously, down regulation of RNA 1 leads to an increase in the abundance of common miRNAs that target RNA 2 transcripts, and consequently RNA 2 down regulation.

During my doctoral studies, I have systematically reverse-engineered: 1) miR context-specific post-transcriptional targets (Chapter 2); 2) the competing endogenous RNA regulatory network in glioblastoma (Chapter 3); and 3) a high-confident pan-cancer ceRNA network (PC-ceRNET) across multiple tumor types (Chapter 4). For interactions in the PC-ceRNET network, they will be applied to explain how genetic and epigenetic alterations are propagated to regulate key drivers of tumorigenesis (Chapter 5). Together with bioinformatics predictions and experimental validation in cell lines, I will offer evidence supporting the conclusion that, even before translation, mRNAs can play important roles in gene regulation and affect disease pathogenesis via ceRNA interactions by targeting tumor-suppressors or oncogenes. My advisor and I conceived this entire dissertation work and my role primarily focused on algorithm design, data and statistical analysis, and software development. Experimental validation were performed by my colleagues. The greatest technical innovation of this dissertation include 1) the high-throughput prediction and wet-lab validation of miR targets with much lower false positive rates than several existing approaches; 2) the first high-throughput attempt to identify mRNA targets of co-regulating miRs, thereby elucidating the mechanisms of synergy or competition of miR target selection; 3) the first large-scale and genome-wide

study to identify ceRNA interactions in glioblastoma as well as in other three tumor types including breast, prostate, and ovarian cancers. The greatest translational innovation of my dissertation is the discovery of regulatory network modules from the pan-caner ceRNA network that amplify and propagate the effects of genetic and epigenetic alterations to drive pathology in a variety of contexts. When considered within an integrative context, these regulatory modules will help identify therapeutic and diagnostic biomarkers, and master regulators of distinctive gene-expression signatures in tumor subtypes. In addition, my dissertation work will improve our understanding of cellular regulation and how the various regulatory layers interact to affect cellular programs. Every specific project in the following chapters are outlined as follows.

1.2 Predicting miR targets

Accurate miR target predictions are desired to improve our understanding of post-transcriptional regulation, but current prediction methods are notoriously inaccurate; see Figure 1.2. To identify high-confidence miR targets, I have developed an integrative approach called *Cupid*, which uses a 3-step process to predict miR-target interactions:

Step1: Cupid scores miR binding sites in 3' UTRs using 1) sequence-based binding-site predictions made by TargetScan, PITA and MIRANDA [21-23], 2) 46-vertebrate genome cross-species conservation scores by PhastCons [24], and 3) positional information relative to the 3' UTR start site. Specifically, Cupid trains an SVM classifier using LIBSVM to produce scores from 0 to 1 for each site that is predicted by at least one of the three algorithms by training against 684 validated miR targets obtained from miRecords [25] as of June, 2010.

Step 2: Site scores are then summarized using an array of summary functions that model miR binding-site interactions both linearly and non-linearly, and an SVM is then used to generate miR-target interaction scores using site scores and their summaries with co-expression data. The level of co-expression between miR and its target is measured by the mutual information. The down-sampling and bagging techniques were integrated into both Step 1 and 2 to achieve a higher precision rate.

Step 3: Cupid predicts functional, context-specific interactions, by evaluating evidence for synergistic regulation between miRs and evidence for mRNA competition for miR regulation, and by inferring post-transcriptional down-regulation of transcription factors and signaling molecules from changes in the expression of their targets inferred by ARACNe [26] or supported by LINCS [27] data. At this step, Cupid

identifies regulators targeted by miRs by correlating miR expression and the expression of the predicted targets of the regulators.

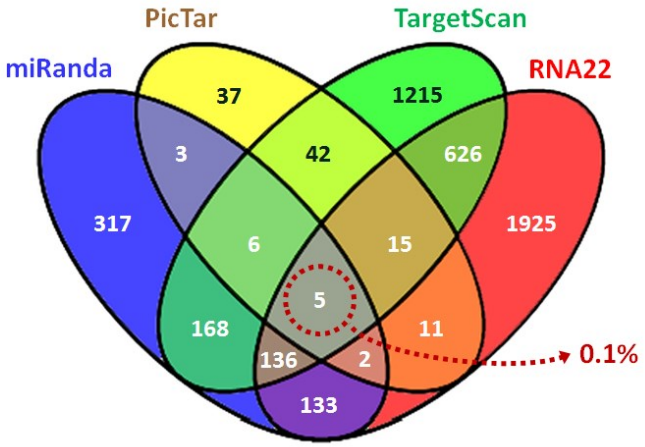


Figure 1.2. Prediction of human miR-28 target sites

A four-way Venn diagram shows there is very poor concordance between different miRNA-binding site prediction methods suggests that additional improvement is desired.

For the experimental analysis, miR target predictions will be validated *en masse* with transfection experiments by individual miRs or co-regulating miRs followed by western blot analysis of their target genes. Luciferase reporter assays with intact miR binding sites downstream of the reporter will be used to demonstrate direct interactions for targets with therapeutic potential. Transfection of competing 3' UTRs, siRNA-mediated silencing, followed by western blot and luciferase assays will be used to demonstrate mRNA competition for miR regulation.

In summary, by using a stringent selection criterion to predict interactions, Cupid identified about 1/2 million candidate interactions, which is fewer than the number of interactions that are common to TargetScan, PITA and MIRANDA predictions. Cupid's evidence integration scheme is complex, and while these three methods agree on 8% of their calls, their common interactions include only 6% of Cupid-predicted interactions. Initial validation suggests a high prediction success rate (80% true positive rate) even when only one site per target 3' UTR is tested. Method comparisons on high-quality miR-target data, including PAR-CLIP [28], miR transfection experiments [29-31] from GEO, and transfection of miRNA mimics with

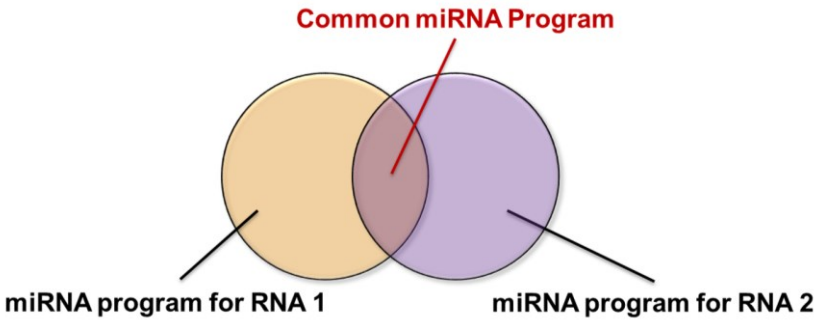
RPPA output suggest superior predictive accuracy. Note that Cupid is specifically designed to produce very low false-positive rates, possibly at the expense of false-negative rates, by using stringent, integrative selection criteria, which can provide a solid foundation for studying ceRNA interactions on a genome scale (see Chapter 3 and 4).

1.3 Reverse engineering ceRNA network in glioblastoma

To evaluate both the range and potential tumorigenic role of ceRNA interactions, I have presented a new multivariate analysis method called *Hermes*. By analyzing a large set of sample-matched gene and miR expression profiles from The Cancer Genome Atlas (TCGA), I have uncovered a posttranscriptional regulation layer of surprising magnitude, comprising more than 248,000 microRNA (miR)-mediated interactions in glioblastoma [14]. These include about 7,000 genes whose transcripts act as miR "sponges". Biochemical analyses confirmed that this network regulates established drivers of tumor type and subtype initiation, including *PTEN*, *PDGFRA*, *RB1*, *VEGFA*, *STAT3*, and *RUNX1*, suggesting that these interactions mediate crosstalk between canonical oncogenic pathways. The ceRNA network provides a mechanistic, experimentally validated rationale for the loss of PTEN expression in a large number of glioma samples. Moreover, in addition to PTEN, I identified nearly 200 genes, including many known drivers of tumorigenesis and tumor subtype, whose expression profiles had stronger correlation with deletions at the loci of their ceRNA regulators than with deletions at their own loci. I believe that ceRNA interactions provide channels for the propagation of genetic alterations to affect distal loci, and may point to the origins of previously unexplained regulation. This regulatory network fills in a missing piece in the puzzle of cell regulation, and will help researchers track down genetic and epigenetic alterations that are propagated by ceRNA interactions to distally affect tumor-specific gene expression programs. In the following, I will briefly describe the methodology of Hermes; see Figure 1.3.

Hermes predicts ceRNA interactions based on the relative size of shared miRNA regulatory programs between two genes based on predictions by the Cupid algorithm (Chapter 2), and the conditional mutual information between these genes and their shared miRNA program. Namely, given genes *T* and *R*, and the set of miRNAs that regulate them $\Pi_{miR}(T)$ and $\Pi_{miR}(R)$, their shared program is identified by taking the intersection $\Pi_{miR}(T,R) = \Pi_{miR}(T) \cap \Pi_{miR}(R)$. First, Hermes tests that the size of $\Pi_{miR}(T,R)$ relative to the sizes of the individual programs is statistically significant at FDR<0.01 by weighted Fisher's exact test. Then,

(A)



(B)

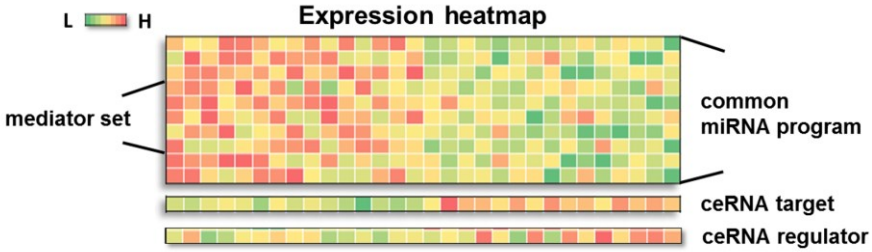


Figure 1.3. Hermes methodology

Hermes is an information theoretic approach used to search for ceRNA regulators by looking for two properties: (A) Significant common miRNA program. For each RNA pair, compute the overlap of two miRNA programs and estimate its significance by weighted Fisher's Exact Test. The known and predicted miRNA targets in 3' UTRs are collected from Cupid. (B) Significant evidence for conditional regulation. Compute conditional mutual information (CMI) to identify the ceRNA regulator (*R*) by finding a gene whose expression is associated with changes in mutual information between the common miRNA program ($\cup miRNA$) and the ceRNA target (*T*).

Hermes evaluates the statistical significance p_i (*p*-value) of the test $I[miR_i; \mathrm{T}|R] > I[miR_i; T]$, where the variables indicate the expression of the corresponding RNA species. The CMI is estimated using an adaptive partitioning algorithm [32] by first iteratively partitioning the 3-dimentional expression space evenly into 8 partitions per iteration until partitions are balanced (p>0.05 by Chi-squared test), and then summing up CMI across partitions. *P*-values for each triplet are computed using a null-hypothesis where the candidate ceRNA regulator's expression (R) is shuffled 1,000 times, thus preserving the pairwise mutual information between miRNA and target. Final significance across the entire program is using weighted

Brown's method to integrate each of regulatory directions, i.e. *R* affecting *miR*$_i$ regulation of *T* as well as *T* affecting *miR*$_i$ regulation of *R*, for all the miRNAs in the shared miRNA program $\Pi_{miR}(T;\mathrm{R})$. Finally, only prediction passing significance of FDR<1E-05 were selected. Note that selected predictions by Hermes have been validated in two glioblastoma cell lines, SNB19 and SF188 [14].

1.4 Constructing the pan-cancer ceRNA network across multiple tumors

Since Hermes requires large-scale, sample-matched gene and miR expression profiles, the existence of a high-confident network will allow us to examine the effects of these interactions across tumor types even for tumors with few profiled patient samples. Here, the goal is to use Cupid and Hermes to construct ceRNETs for glioblastoma using gene and miRNA expression [33] (423 samples, 12,032 genes, 469 miRNAs profiled), ovarian cancer [34] (583 samples, 12,032 genes, 713 miRNAs profiled), prostate cancer [35] (140 samples, 23,614 genes, 367 miRNAs profiled) and breast cancer [36] (207 samples, 18,748 genes, 524 miRNAs profiled). The resulting predicted ceRNETs had 527,430 (glioblastoma), 532,869 (ovarian), 476,456 (prostate) and 447,011 (breast) predicted interactions. The pan-cancer ceRNA network includes 164,623 ceRNA interactions after taking the intersection of four ceRNETs, about a third of those found in each individual tumor context at FDR<1E-05, were predicted to be ubiquitous across all four ceRNETs.

To test the statistical significance of this overlap, I performed permutation tests, followed by overlap analysis. In each test, candidate ceRNAs and their individual number of interactions were fixed to those inferred in each tumor type, and edge swapping was restricted to candidate interactions with significantly many candidate miRNA regulators. In total, despite performing 10^{12} permutation tests, I never observed an overlap of size that is comparable to the one obtained from my inferred networks, suggesting a high statistical significance (p < 1E-12) for the size of this overlap. In the following, I will refer to the high-confident subnetwork that is common to all four tumors types as the pan-cancer ceRNA network (*PC-ceRNET*).

In order to identify miRNA *mediators* in addition to ceRNA interactions, I modified the Hermes algorithm to predict miRNA mediators and to account for miRNA-target binding scores and co-expression between miRNA species; predicted miRNA mediators of each candidate interaction (*T*, *R*) are the set of miRNAs that T and R are predicted to compete for; see Chapter 4 for details. The resulting networks suggest that while almost all ceRNA interactions are implemented by miR programs with tumor type-specific expression,

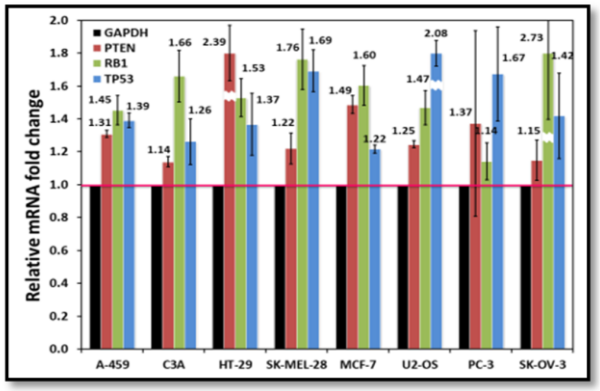
hundreds of thousands of interactions that regulate gene expression through these context-specific implementations are themselves context independent. To further quantify this finding, I computed the Krippendorff's alpha (*α*) coefficient [37-38] for each ceRNA interaction in the PC-ceRNET. For each conserved interaction, these coefficients describe the magnitude of the overlap between Hermes-predicted miRNA mediators across the four networks, and are compared to a null distribution obtained from bootstrapping.

Lastly, a unique property of ceRNETs is an increase in correlation between the expression of a specific ceRNA target and the total expression of its ceRNA regulators, as a function of the number of regulators, which has been attributed to combinatorial regulation by ceRNAs [8]. To evaluate the predictive power of PC-ceRNET interactions, I reported (1) an evaluation of the predictive ability of the PC-ceRNET on the expression of ceRNA targets in both tumor-related and non-tumor context, and (2) median correlations between ceRNA-target expression profiles and the standardized totals of the expression profiles of their predicted regulators. I will use a ridge-regression with Glmnet for Matlab within a 10-fold cross validation analysis scheme [39-40] to predict the expression of each PC-ceRNET target from the expression of its inferred ceRNA regulators. For each ceRNA target, in each 10-fold cross validation step, Glmnet constructs a regression model using training samples to fit an estimate $\hat{y}$ for ceRNA-target expression testing-sample profile y. The test-set residuals ($\hat{\varepsilon}$) are then compiled across the 10 testing-sample sets by taking the difference between the ceRNA-target expression profile y and the fitted estimate $\hat{y}$, so that $\hat{\varepsilon} = y - \hat{y}$. To calculate R^2, I take the sum of the square of the residuals across all samples, $R^2 = 1 - \sum_i \hat{\varepsilon}_i^2 \,/\, \sum_i (y_i - \bar{y})^2$, where $\bar{y}$ is the mean expression of the ceRNA target across the dataset. To assign *p*-values, to the predictive ability I used bootstrapping. Namely, the ceRNA-target expression profile y is adjusted so that $y' = \hat{y} + \delta$, where $|\delta| = |\hat{\varepsilon}|$ and δ is populated by random selection from $\hat{\varepsilon}$, with replacement. The Glmnet regression was repeated for one thousand bootstrapping y's, estimating bootstrapping R^2 using 10-fold cross validation analysis to produce a null distribution.

For the experimental validation, I will provide evidence for PC-ceRNET interactions by transfection with 3' UTR of ceRNA regulators followed by qPCR to measure changes in target expression profiles. The siRNA-mediated silencing of *DICER* and *DROSHA*, which are necessary for miR processing, will be performed to investigate if the effect of ceRNA interactions will be abrogated sufficiently, implying these interactions are

miR mediated. A sample experimental results, shown in Figure 1.4, suggest that ceRNA interactions between three established tumor suppressors, PTEN, RB1, and TP53, are ubiquitous in a variety of human cancer cell lines.

(A)



(B)

UTRs / Genes	PTEN	RB1	TP53
PTEN	7|7	6|7	6|7
RB1	6|8	7|8	7|8
TP53	5|8	4|8	7|8

Figure 1.4. Sample experimental results for demonstrating ceRNA interactions are ubiquitous

(A) Transfections of *TP53* 3' UTR increased expression of tumor suppressors in different tumor cells. Other transfection experiments are not shown. (B) The number of cell lines were the transfection of a UTR (column) lead to significant (p<0.01) increase to the expression of a profiled gene (row) in cell lines were this gene is expressed.

1.5. Identifying PC-ceRNA interactions that account for missing genetic or epigenetic variability in tumors

Despite our efforts to use next generation sequencing to identify genetic and epigenetic factors that drive tumorigenesis, driver alternations for many patients remain unknown [41-43]. A considerable research effort has therefore been undertaken to discover new factors which are likely to involve in cancer development [44-47]. I and my colleagues have previously shown that, using simple correlation analysis, that downregulation of the tumor suppressor PTEN in glioblastoma can be predicted through copy-number deletions of the loci of its ceRNA regulators when the PTEN locus is intact, suggesting that ceRNA interactions are capable of explaining missing genetic variability [14]. Yet, a more sophisticated and systematic approach is required to study how the combination of genetic (copy number variation) and

epigenetic (DNA methylation) alterations in ceRNA regulators affect the expression of key drivers of tumorigenesis. Based on the previous aim, however, the PC-ceRNET provides a causal framework for identifying genomic alterations in ceRNAs that may cooperatively dysregulate the expression of specific genes of interest.

I focus on eight tumor datasets, including glioblastoma, as well as carcinomas of the colon, head and neck, kidney, ovary, uterus, prostate, and breast, my analysis uncovered a large repertoire of ceRNAs whose genomic alterations may contribute to dysregulation of thousands of genes, including a large number of established cancer genes. First, I used elastic-net regression [33,37] with 10-fold cross validation to identify *ceRNA drivers*, whose expression is significantly predictive of the aberrant expression of genes with missing genomic variability. In total, ceRNA drivers were identified for >85% of PC-ceRNET genes missing genomic variability, across eight tumor contexts. The PC-ceRNET (Chapter 4) will be used as a universal tool to study each tumor independently.

Specifically, given expression profiles for a ceRNA target and its *N* predicted ceRNA regulators, I selected candidate drivers by first clustering them according to their expression profiles. Clustering was performed using *k*-means with all possible choices for *k*, where each cluster is represented by its centroid. Then, for each *k*, elastic net regression and 10-fold cross validation was used to estimate a test-set residual sum of squares and the corresponding Akaike information criterion (AIC) [48]. Note that elastic net regression is commonly used for identifying interactions [49]. AIC is a distance measure that punishes likelihood functions that are based on more variables, and 10-fold cross validation was used here to rank and select solutions rather than evaluate their overall significance. Genes contributing to at least 50% of the top $\sqrt{N}$ results by AIC, after sample size correction, were selected as candidate drivers. Finally, I summed across standardized expression profiles of candidate drivers and compared the correlation between these total profiles and the expression profile of the target gene. To assign significance for this final selection of ceRNAs, I compared the resulting correlation coefficient to a distribution of correlations obtained by shuffling sample labels (selecting $p < 0.05$).

I clustered ceRNA regulators and represented them by cluster centroids (super genes) to improve prediction rates and aid in significance testing [50], while allowing for the inclusion of correlated ceRNA regulators during candidate driver selection following regression. Specifically, elastic net regression produces

regression models with sparse variable selections. By representing correlated genes as aggregate variables I reduce the number of variables for selection by elastic net regression while ensuring that correlated drivers, which may be omitted by elastic net regression because their simultaneous inclusion in a predictive model does not improve the fit, could be considered when making candidate driver selections. Thus, after centroid selection by elastic net regression all represented ceRNA regulators are considered in the next selection step.

Similarly to the procedure described for evaluating the predictive power of PC-ceRNET interactions, regression with 10-fold cross validation was used to estimate the reduction in variance. Then Akaike information criterion (AIC) was computed as $\mathrm{AIC} = n \ln(\sum_i \hat{\varepsilon}_i^2 / n) + 2k$, where *n* is the number of samples in the dataset and *k* is the number of clusters used. To correct for small sample size and avoid overfitting the data, especially in cases where n is small relative to k, I used the sample-size correction $\mathrm{AIC}' = \mathrm{AIC} + \frac{2k(k+1)}{n-k-1}$ [51]. This criterion was used to compare regression models across *k*, where lower AIC' is associated with improved sample-size corrected predictive power. Expanding centroid to genes they represent, selecting as candidate drivers those ceRNA regulators that that contributed to at least 50% of the top $\sqrt{N}$ results by AIC'.

References

1. Bartel, D.P. MicroRNAs: target recognition and regulatory functions. *Cell* 136, 215-233 (2009).

2. Brodersen, P. & Voinnet, O. Revisiting the principles of microRNA target recognition and mode of action. *Nature reviews. Molecular cell biology* 10, 141-148 (2009).

3. Gabriely, G. et al. Human glioma growth is controlled by microRNA-10b. *Cancer research* 71, 3563-3572 (2011).

4. Godlewski, J. et al. Targeting of the Bmi-1 oncogene/stem cell renewal factor by microRNA-128 inhibits glioma proliferation and self-renewal. *Cancer research* 68, 9125-9130 (2008).

5. Kim, M.S. et al. Somatic mutations and losses of expression of microRNA regulation-related genes AGO2 and TNRC6A in gastric and colorectal cancers. *The Journal of pathology* 221, 139-146 (2010).

6. Kwak, H.J. et al. Downregulation of Spry2 by miR-21 triggers malignancy in human gliomas. *Oncogene* 30, 2433-2442 (2011).

7. Lu, J. et al. MicroRNA expression profiles classify human cancers. *Nature* 435, 834-838 (2005).

8. Cancer Genome Atlas Research, N. Integrated genomic analyses of ovarian carcinoma. *Nature* 474, 609-615 (2011).

9. Kim, T.M., Huang, W., Park, R., Park, P.J. & Johnson, M.D. A developmental taxonomy of glioblastoma defined and maintained by MicroRNAs. *Cancer research* 71, 3387-3399 (2011).

10. Poliseno, L. et al. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* 465, 1033-1038 (2010).

11. Salmena, L., Poliseno, L., Tay, Y., Kats, L. & Pandolfi, P.P. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* 146, 353-358 (2011).

12. Tay, Y. et al. Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. *Cell* 147, 344-357 (2011).

13. Karreth, F.A. et al. In vivo identification of tumor- suppressive PTEN ceRNAs in an oncogenic BRAF-induced mouse model of melanoma. *Cell* 147, 382-395 (2011).

14. Sumazin, P. et al. An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell* 147, 370-381 (2011).

15. Marques, A.C., Tan, J. & Ponting, C.P. Wrangling for microRNAs provokes much crosstalk. *Genome biology* 12, 132 (2011).

16. Franco-Zorrilla, J.M. et al. Target mimicry provides a new mechanism for regulation of microRNA activity. *Nature genetics* 39, 1033-1037 (2007).

17. Lee, D.Y. et al. Expression of versican 3'-untranslated region modulates endogenous microRNA functions. *PloS one* 5, e13599 (2010).

18. Cazalla, D., Yario, T. & Steitz, J.A. Down-regulation of a host microRNA by a Herpesvirus saimiri noncoding RNA. *Science* 328, 1563-1566 (2010).

19. Cesana, M. et al. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* 147, 358-369 (2011).

20. Ebert, M.S. & Sharp, P.A. Roles for microRNAs in conferring robustness to biological processes. *Cell* 149, 515-524 (2012).

21. Lewis, B.P., Burge, C.B. & Bartel, D.P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120, 15-20 (2005).

22. Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U. & Segal, E. The role of site accessibility in microRNA target recognition. *Nature genetics* 39, 1278-1284 (2007).

23. John, B. et al. Human MicroRNA targets. *PLoS biology* 2, e363 (2004).

24. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research* 15, 1034-1050 (2005).

25. Xiao, F. et al. miRecords: an integrated resource for microRNA-target interactions. *Nucleic acids research* 37, D105-110 (2009).

26. Margolin, A.A. et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics* 7 Suppl 1, S7 (2006).

27. Peck, D. et al. A method for high-throughput gene expression signature analysis. *Genome biology* 7, R61 (2006).

28. Hafner, M. et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* 141, 129-141 (2010).

29. Leivonen, S.K. et al. Protein lysate microarray analysis to identify microRNAs regulating estrogen receptor signaling in breast cancer cell lines. *Oncogene* 28, 3926-3936 (2009).

30. Gotte, M. et al. miR-145-dependent targeting of junctional adhesion molecule A and modulation of fascin expression are associated with reduced breast cancer cell motility and invasiveness. *Oncogene* 29, 6569-6580 (2010).

31. Frankel, L.B. et al. microRNA-101 is a potent inhibitor of autophagy. *The EMBO journal* 30, 4628-4641 (2011).

32. Darbellay , G. and I. Vajda, *Estimation of the Information by an Adaptive Partitioning of the Observation Space.* IEEE Trans. on Information Theory, 1999. 45: p. 1315--1321.

33. Cancer Genome Atlas Research, N. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061-1068 (2008).

34. Cancer Genome Atlas Research, N. Integrated genomic analyses of ovarian carcinoma. *Nature* 474, 609-615 (2011).

35. Taylor, B.S. et al. Integrative genomic profiling of human prostate cancer. *Cancer cell* 18, 11-22 (2010).

36. Buffa, F.M. et al. microRNA-associated progression pathways and potential therapeutic targets identified by integrated mRNA and microRNA expression profiling in breast cancer. *Cancer research* 71, 5635-5645 (2011).

37. Krippendorff, K., *Estimating the reliability, systematic error, and random error of interval data.* Educational and Psychological Measurement, 1970. 30 (1): p. 61-70.

38. Krippendorff, K., *Content analysis : an introduction to its methodology*. 2nd ed. 2004, Thousand Oaks, Calif.: Sage. xxiii, 413 p.

39. Zou, H. and T. Hastie, *Regularization and variable selection via the elastic net.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005. 67(2): p. 301-320.

40. Tibshirani, R., *Regression shrinkage and selection via the lasso.* J. Royal. Statist. Soc B, 1996. 58(1): p. 267-288.

41. Pleasance, E.D., Cheetham, R.K., Stephens, P.J., McBride, D.J., Humphray, S.J., Greenman, C.D., Varela, I., Lin, M.L., Ordonez, G.R., Bignell, G.R.*, et al.* (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. Nature *463*, 191-196.

42. Meyerson, M., Gabriel, S., and Getz, G. (2010). Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet *11*, 685-696.

43. Negrini, S., Gorgoulis, V.G., and Halazonetis, T.D. (2010). Genomic instability--an evolving hallmark of cancer. Nat Rev Mol Cell Biol *11*, 220-228.

44. Takata, R., Akamatsu, S., Kubo, M., Takahashi, A., Hosono, N., Kawaguchi, T., Tsunoda, T., Inazawa, J., Kamatani, N., Ogawa, O.*, et al.* (2010). Genome-wide association study identifies five new susceptibility loci for prostate cancer in the Japanese population. Nat Genet *42*, 751-754.

45. Santarius, T., Shipley, J., Brewer, D., Stratton, M.R., and Cooper, C.S. (2010). A census of amplified and overexpressed human cancer genes. Nat Rev Cancer *10*, 59-64.

46. Gupta, R.A., Shah, N., Wang, K.C., Kim, J., Horlings, H.M., Wong, D.J., Tsai, M.C., Hung, T., Argani, P., Rinn, J.L.*, et al.* (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature *464*, 1071-1076.

47. Wu, W.K., Lee, C.W., Cho, C.H., Fan, D., Wu, K., Yu, J., and Sung, J.J. (2010). MicroRNA dysregulation in gastric cancer: a new player enters the game. Oncogene *29*, 5761-5771.

48. Akaike, H., *A new look at the statistical model identification.* IEEE Transactions on Automatic Control, 1974. 19(6): p. 716–723.

49. Barretina, J., et al., *The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.* Nature, 2012. 483(7391): p. 603-7.

50. Park, M.Y., T. Hastie, and R. Tibshirani, *Averaged gene expressions for regression.* Biostatistics, 2007. 8(2): p. 212-27.

51. Burnham, K.P., D.R. Anderson, and K.P. Burnham, *Model selection and multimodel inference : a practical information-theoretic approach*. 2nd ed. 2002, New York: Springer. xxvi, 488 p.

Chapter 2: Cupid – an integrative approach for miRNA target prediction

2.1 Introduction

MicroRNAs regulate gene expression by modulating target RNA stability and translation [1]. Their dysregulation has been implicated in a wide range of human diseases including cancer [2]. The effects of miRNA regulation are context specific and depend on their tissue-specific abundance [3] as well as the abundance and localization of their targets [4,5]. Knowledge about their functional targets, in a given context, is a necessary step towards understanding their effects on cellular behavior and disease. High-throughput methods to profile miRNA-target interactions on genome-wide scales include HITS-CLIP [6], PAR-CLIP [7], and CLASH [8], but these cannot be used to identify miRNA-target interactions in disease samples. Thus, computational prediction methods are necessary for systematic identification of candidate miRNA-target interactions that influence disease.

The earliest computational miRNA-target prediction methods, including miRanda [9] and RNA22 [10], were based on sequence alignments of mature miRNAs and their candidate targets. These methods have high false discovery rates, and biochemical experiments suggest that even high-confidence sequence-alignment based interaction predictions may not be functional. Additional constraints were needed to reduce the number of false positive calls. Some of the earliest methods, including miRanda [9] and TargetScan [11] used cross-species conservation as a feature to identify likely functional miRNA-binding regions [12]. Others incorporate RNA-expression based evidence to address context specificity of miRNA-target regulation [13-15]. However, both cross-species conservation and miRNA-target anti-correlation are weak predictive features of functional regulation by miRNAs [16,17], and an optimization process should be used when integrating them within a predictive framework.

I use systems-biology approach to infer functional context-specific regulation by miRNAs, considering evidence that putative miRNA targets compete for regulation by their common targeting miRNAs, evidence for synergy [18,19] between miRNA species that are predicted to share targets, and evidence for indirect regulation by miRNAs. I have demonstrated that integrating evidence for functional regulation with more traditional evidence for miRNA targeting strikes an improved tradeoff between precision and recall through substantial improvement in precision. My method, Cupid, first predicts miRNA binding sites in 3' UTRs of

candidate targets based on sequence alignments and scores sites by comparing them to previously validated predictions. It then integrates predicted binding-site attributes with multivariate co-expression to predict interactions. Finally, Cupid tests predicted interactions for evidence for functional regulation.

Cupid was used to predict functional targets of miRNA regulation in breast cancer tumors, and my colleagues provide both low-throughput and high-throughput validation for its predictions in breast cancer cell lines. Evidence for regulation includes microarray-based gene expression profiling following miRNA-precursor transfections, protein expression profiling using 158 antibodies following mimic transfections for 159 miRNAs, and 3' UTR luciferase activity assays following transfection of miRNA mimics. Computational evidence suggests that CCND1, ESR1, HIF1A and PDGFRA and NCOA3, which have been previously implicated in regulating breast cancer tumorigenesis, compete for regulation by miRNAs. Biochemical assays in MCF7, a breast cancer cell line, demonstrate this regulatory potential. Analysis of breast cancer tumors point to ten miRNAs that potentially regulate these genes, and luciferase activity assays support the regulation of CCND1, ESR1, HIF1A and PDGFRA by each of miR-17-5p, miR-18a/b-5p, miR-106b-5p, miR-130a/b-3p and miR-301a-3p. Using miRNA perturbations followed by protein-expression profiling in MDA-MB-231, I provide biochemical evidence for over 230 potential miRNA-target interactions. I also observed significant regulatory potential by multiple miRNAs for 30 genes, including AKT2, AKT3, CCND1, FOXO3, IRS1, MAPK1, MAP2K1, MET, MYC, NFKB1, PIK3CA, RB1, SMAD1 and SMAD3. Moreover, computational analyses of RPPA-derived protein expression profiles in breast cancer tumors, point to significant regulation of ESR1 by over forty miRNAs, and my colleagues verified the regulatory potential of twelve of these using ESR1 3' UTR luciferase activity assays. In addition, through analysis of gene expression experiments following miRNA perturbations in MCF7 [20,21] and MDA-MB-231 [22], I provide expression based evidence for 206 miRNA-target interactions in breast cancer tumors.

In the following sessions, I begin by describing the methodology underlying Cupid, my proposed miRNA-target prediction method. Cupid proceeds in three phases, as depicted in Figure. 2.1(A). I then describe my efforts to test predicted sites and interactions, including testing of predicted binding sites using PAR-CLIP data, and testing predicted interactions and predicted functional interactions in perturbation experiments in breast cancer cell lines and across RNA and protein expression profiles in breast cancer tumors. I describe

my attempt to identify functional evidence for predicted interactions through indirect regulation detected at the level of the targets of genes (effectors) that are predicted to be targeted by the miRNA (Figure. 2.1(B)).

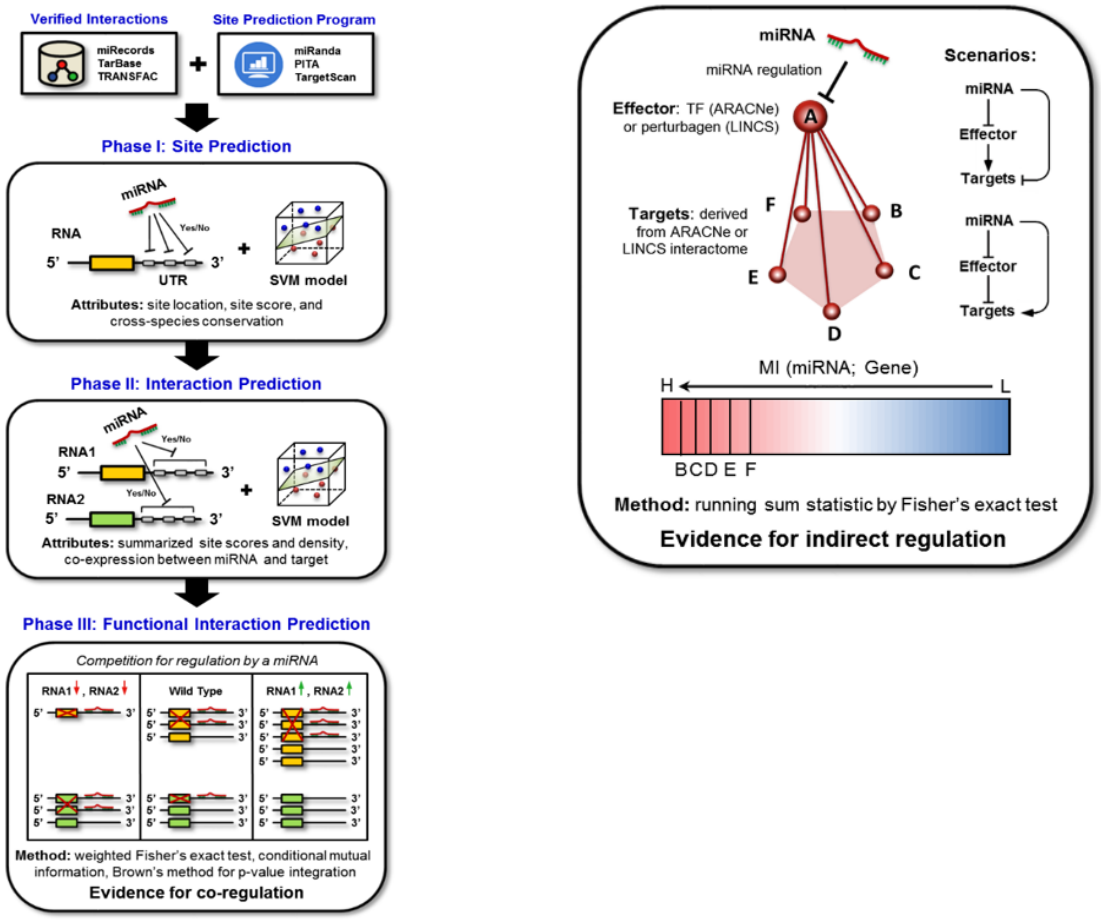


Figure 2.1. Cupid methodology

(A) Cupid proceeds by first reevaluating sites predicted by TargetScan, miRanda and PITA (Phase I). It selects sites and produces probabilistic site scores for each candidate site after considering its location, estimated quality and degree of cross-species conservation, and after comparing these predictive features to those of verified interactions. Selected sites for each miRNA and candidate target 3' UTR are used to select and probabilistically score miRNA-target interactions (Phase II), which are then examined for evidence for functional regulation (Phase III). I use evidence for competition for miRNA regulation in breast cancer tumors as evidence for functional regulation. In addition, I also consider (B) evidence for indirect regulation by miRNAs as functional evidence for regulation by miRNAs.

I use the term 'effectors' to highlight the role of miRNA target genes as intermediaries that channel the effects of miRNA regulation to sets of downstream genes. Finally, focusing on predicted miRNA regulators

of ESR1, I show how multiple lines of evidence arising from genome-wide profiling data and systems-level regulatory network analyses can be integrated to produce accurate context-specific predictions.

2.2 Cupid

Cupid predicts functional miRNA-target interactions in three phases. First, Cupid rescores sites predicted by TargetScan [23], miRanda [24] and PITA [25] by comparing their scores, as provided by the methods that predicted them, their location in the 3' UTR, and their degree of cross-species conservation to these same predictive features for previously validated and high-confidence sites. Then, in the second phase, it predicts miRNA-target interactions by evaluating selected sites, their multiplicity and the multivariate correlation between the expression profiles of the miRNA and its putative target. The evaluation process includes comparing these predictive features to those of previously validated interactions. Finally, in the third phase, Cupid assess whether predicted interactions may be functional in a given context using evidence that putative miRNA targets compete for regulation by their common targeting miRNA. Here I report on prediction in TCGA breast cancer samples [26], and describe Cupid with details relevant to this dataset.

2.2.1 Framework

Cupid uses previously identified sites to guide an SVM-based learning process and predict new sites. A total of 588 previously identified miRNA-RefSeq target interactions, corresponding to 1481 putative or verified sites were collected from TarBase (TarBase_V5) [27], TRANSFAC (Release 2009.3, October 2009) [28] and miRecords (March 2010) [29].

Binding site predictions for 1,218 miRNAs in miRBase (Release 16) [30] by TargetScan [23], miRanda [24] and PITA [25] in 20,491 RefSeq 3' UTRs, associated with 18,093 genes. Overlapping predicted sites from multiple prediction methods (overlaps of one base or more) were attributed to all contributing prediction methods. In total, 36,986,648 sites, corresponding to 11,542,856 interactions, were predicted in RefSeq 3' UTRs, with no evidence from curated literature. Prediction scores were quintile normalized to produce scores in [0, 1].

When predicting both sites and interactions, I used LIBSVM [31] to score candidates. Given that the number of candidate interactions dwarfs the number of previously identified interactions, down sampling was

required to effectively distinguish between candidates with similar properties to those previously identified. When predicting sites, I randomly sampled 1% of candidates (370K sites and 115K interactions) and proceeded to cluster them according to their predictive properties. When building SVM classifiers, clusters were represented by site and interaction representatives that were chosen by chance. Ten-fold cross validation was ran using these representatives and previously identified sites or interactions to select a (cost, γ) combination for a final classifier that was used to score all candidates. To fine tune parameter selection, accuracy maximization, evaluated using a Radial Basis Function kernel, was performed using a grid search process. Probability estimates are a confidence measure for the classification using the final classifier [32], trained on all cluster representatives and using the optimal (cost, γ) combination.

This process was repeated 1,000 times, resampling, clustering and selecting representatives *de novo* at each run to produce 1,000 inclusion probabilities and decisions for each candidate. Finally, each candidate was scored based on the number of inclusion decisions (known as bagging) and average of the probabilities across these bootstrapping runs.

2.2.2 Site predictions

Sites were predicted and scored by TargetScan, miRanda and PITA using default parameters in RefSeq-defined 3' UTRs on December 3rd, 2010, which include 20,491 transcripts for 18,093 genes. In total, the three algorithms predicted 37M distinct sites; see Figure. 2.2(A). Individual site scores were quantile normalized, and sites were assigned a normalized distance from the start of the 3' UTR. Sites were tested for cross-species conservation by requiring binding-site seeds, aligned to miRNA position 2 to 8, to maintain a geometric average per-position conservation probability greater than 95%, according to PhastCons based on 46 vertebrate genomes [33]. Site features were compared to features of previously validated sites or sites in 3' UTRs of previously validated targets using a support vector machine [31]. To do so efficiently, sites were first clustered using K-means into 1481 clusters, matching the number of sites representing validated interactions. Each cluster was represented by at least one randomly selected site, and large clusters were proportionally represented so that ten representatives were selected for a cluster that is ten time the size of the smallest cluster. A classifier was trained on the selected representatives within a 10-fold cross validation framework, producing a test probability and an exclusion/inclusion decision for each binding-site candidate. The process was repeated one thousand times with representative sets chosen *de*

novo at each run, producing an inclusion probability and an inclusion decision for each candidate biding site in each run. Binding site selection was based on a majority vote amongst the 1000 inclusion decisions, and binding-site score was set to be the average probability across runs; see Figure 2.3. In total, almost 1.6 million sites had a consensus inclusion decision (Figure. 2.2(A)).

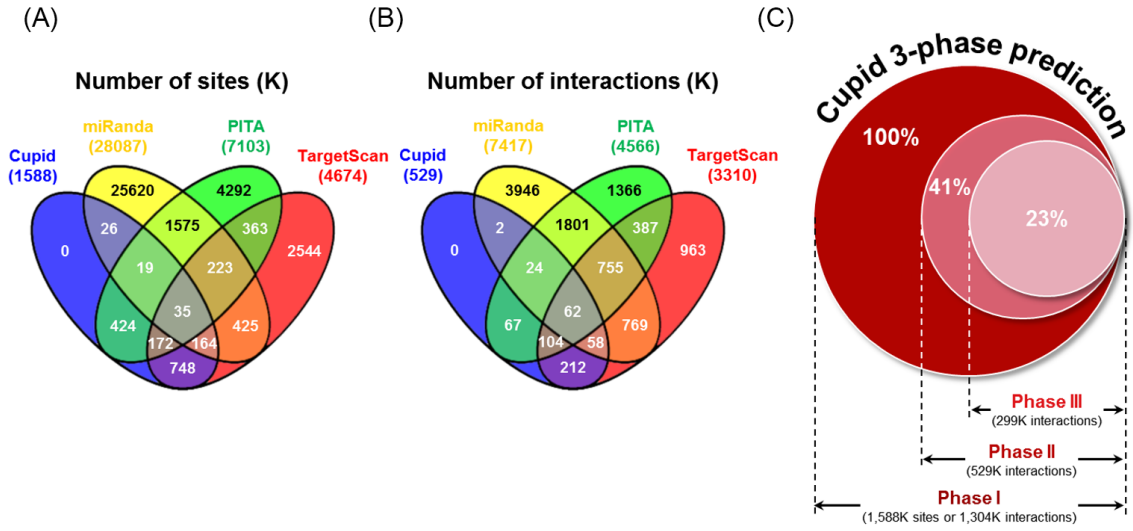


Figure 2.2. Cupid 3-phase prediction

(A) The majority of site predictions by TargetScan, miRanda and PITA are exclusive to a single algorithm; for example, miRanda predicts over 26M sites but only 1M is predicted by TargetScan or PITA. Most Cupid predicted sites were predicted by TargetScan. (B) Cupid predicted 529K miRNA-target interactions in Phase II, excluding 60% of possible interactions. As a result, it makes considerably fewer predictions than TargetScan, miRanda and PITA. (C) Less than a quarter of candidate interactions selected in Phase I were included in Phase III.

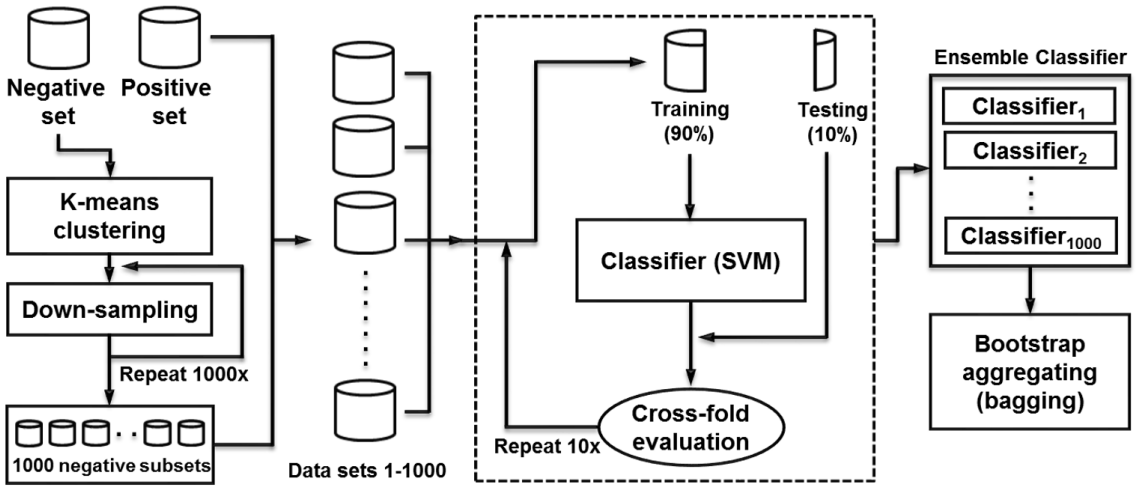


Figure 2.3. Learning site and interaction features

Cupid selects sites and produces probabilistic site scores for each candidate site and interaction after comparing predictive features of candidates to those of verified interactions. The process begins with sampling 1% of candidates and clustering them according to the number of verified interactions. An SVM is then trained on cluster representatives together with validated interactions within a 10-fold cross validation framework to produce probabilistic scores for each candidate interaction. The process is repeated 1000 times and candidates are scored through consensus decisions across bootstrap runs.

2.2.3 Interaction predictions

All 37M candidate binding sites were used to predict miRNA-3' UTR interactions. For each candidate interaction with multiple predicted binding sites, site count, density, distances, and scores were summarized; these classification features were trivialized for interactions with a single binding-site candidate. See below for a complete list of all features used. Note that all features were normalized to [0, 1] to simplify candidate clustering.

- Maximum site score
- Median site score
- Medium range site score (max+min)/2
- Sum of site scores
- Product of sites scores, taken as [1-(1-S1)*(1-S2)*...*(1-Sn)]
- Average of sites scores
- Geometric mean of site scores
- Harmonic mean of site scores

- Root mean square of site scores
- Average sum of squares of site scores
- Weighted mean of site scores, where weights are proportional to the minimum distance from start and end of the 3' UTR
- Sum of site-score squares
- Sum of natural logs of site scores
- Sum of natural exponents of site scores
- Average of site-score squares
- Average of the natural logs of site scores
- Average of the natural exponents of site scores
- The number of sites
- The genomic distance from the most upstream to the most downstream site
- The genomic distance between the closest sites
- The genomic distance between the furthest adjacent sites
- The average distance between adjacent sites

In addition to sequence-based features, candidate interactions were evaluated for context-specific correlation between miRNA and candidate-target expression profiles using normalized mutual information (NMI) as estimated by adaptive partitioning [34]. The normalize mutual information between expression profiles of miRNA *M* and gene *G* was computed as

$$NMI(M,G) = 2 \times \frac{I(M;G)}{H(M) + H(G)}$$

where $I(M;G)$ is the mutual information between the expression profiles of the miRNA and its target gene and $H(M)$ is the entropy of the expression profile of the miRNA [35]. Mutual information was calculated using TCGA breast cancer expression profiles, performed by Illumina sequencing (miRNA-Seq and RNA-Seq), for 728 samples, on July 2012. In total, 1,921 miRNAs and 20,475 genes were profiled.

Predictive features where then compared to the features of 588 previously validated interactions using the framework described above. In total, 529K interactions with score greater than 0.5, a majority vote, were selected (Figure 2.2(B)). Only 0.4% of selected interactions failed to include at least one selected binding

site. For these, high multiplicity of low likelihood sites compensated for the absence of high-likelihood individual sites.

2.2.4 Prediction of functional interactions

Evaluated interactions were tested for evidence that putative miRNA targets compete for regulation by these miRNAs. In order to focus on miRNAs that are titrated through competition between targets, I modified Hermes [36] (see Chapter 3) to predict miRNA *mediators* and to account for miRNA-target binding scores and co-expression between miRNA species; predicted miRNA mediators of each candidate interaction (T_i, T_j) are the set of miRNAs that T_i and T_j are predicted to compete for. Evidence for competition for miRNA regulation was collected by constructing a genome-level network of miRNA-mediated interactions using modified Hermes [36], where each directed interaction between two competing miRNA targets, T_i regulates T_j or $T_i \rightarrow T_j$, that is mediated by miRNAs $\{miR\}$ provides evidence for regulation of T_i and T_j by miRNAs in $\{miR\}$. Below I describe the construction of this network. The construction focuses on evaluating candidate gene (target, regulator) pairs, first identifying candidate interactions between genes T_i and T_j that share a substantial miRNA regulatory program, then identifying a potential set of miRNA mediators $\{miR\}$ for $T_i \rightarrow T_j$ and finally evaluating expression-based evidence that the candidate regulator T_i affects the regulatory potential of $\{miR\}$ on target T_j and vice versa. Correct identification of $\{miR\}$ requires consideration for the binding probabilities between each miRNA and the two genes, and accurate significance estimation requires resolving dependencies between miRNA expression profiles. Below I will describe how Hermes [36] was modified to achieve these goals.

First, I used a weighted Fisher's exact test [37] to evaluate candidate gene pairs that potentially compete for miRNA regulation. The test is based on Cupid interaction scores for each Cupid-evaluated miRNA (1218 miRNAs in total) and each of the two candidate targets. Cupid interaction scores S_i^m and S_j^m for miRNA M_m and targets T_i and T_j are derived from SVM inclusion decisions and range from 0 to 1. The total score over all 1,218 miRNAs is given by the following 2×2 contingency table.

	T_j is a target	T_j is not a target	
T_i is a target	$\left\lceil \sum_m S_i^m S_j^m \right\rceil$	$\left\lceil \sum_m S_i^m (1 - S_j^m) \right\rceil$	
T_i is not a target	$\left\lceil \sum_m (1 - S_i^m) S_j^m \right\rceil$	$\left\lceil \sum_m (1 - S_i^m)(1 - S_j^m) \right\rceil$	

Apply Fisher's exact test on the table above to obtain a p-value estimate for the likelihood of the interaction between T_i and T_j. P values were calculated for all gene pairs, and corrected by FDR using qvality [38]. Candidates with estimated q-value below 1E-02 were included.

Second, I evaluate the statistical significance $p_{i\to j}^m$ (*p*-value) of the test $I[M_m; T_j | T_i] > I[M_m; T_j]$, where the variables indicate the expression of the corresponding RNA species and T_i is a candidate regulator of T_j. The CMI is estimated using an adaptive partitioning algorithm [34] by first iteratively partitioning the 3-dimentional expression space evenly into 8 partitions per iteration until partitions are balanced (p>0.05 by chi-squared test), and then summing up CMI across partitions. P values for each triplet, $P_{i\to j}^m$, are computed using a null-hypothesis where the candidate regulator's expression (T_i) is shuffled 1,000 times, thus preserving the pairwise mutual information between miRNA and target. Candidate miRNA mediators with score $P_{i\to j}^{m\sqrt{S_i^m S_j^m}} \leq 0.002$ were selected.

Lastly, each candidate interaction $T_i \to T_j$, where T_j is proposed to be regulated by T_i through competition for candidate miRNA mediators $E \in \{miR\}$ were then evaluated for expression-profile evidence by combining evidence for all miRNAs in E. Define weighted p-value product $Q_{i\to j} = \prod_{M_m \in E} P_{i\to j}^{m\sqrt{S_i^m S_j^m}}$, then the distribution of $X = -2\log Q_{i\to j}$ can be approximated by $c\chi_f^2$, where χ_f^2 is a chi-square variate with f degrees of freedom, characterized below, without requiring independence of CMI values across miRNAs [39-40].

$$E(X) = 2 \sum_{M_m \in E} \sqrt{S_i^m S_j^m}$$

$$\sigma^2(X) = 4 \sum_{M_m \in E} S_i^m S_j^m + 2 \sum \sum_{m<n} \sqrt{S_i^m S_j^m} \sqrt{S_i^n S_j^n} cov(-2logP_{i\to j}^m, -2logP_{i\to j}^n)$$

$$c = \frac{\sigma^2(X)}{2E(X)} = \frac{2\sum_{M_m \in E} S_i^m S_j^m + \sum\sum_{m<n} \sqrt{S_i^m S_j^m}\sqrt{S_i^n S_j^n} cov(-2logP_{i\to j}^m, -2logP_{i\to j}^n)}{2\sum_{M_m \in E}\sqrt{S_i^m S_j^m}}$$

$$f = \frac{2[E(X)]^2}{\sigma^2(X)} = \frac{4\left(\sum_{M_m \in E}\sqrt{S_i^m S_j^m}\right)^2}{2\sum_{M_m \in E} S_i^m S_j^m + \sum\sum_{m<n}\sqrt{S_i^m S_j^m}\sqrt{S_i^n S_j^n} cov(-2logP_{i\to j}^m, -2logP_{i\to j}^n)}$$

The covariance matrix was numerically estimated using Brown's method [41], see below.

$$cov\left(-2logP_{i\to j}^m, -2logP_{i\to j}^n\right) = \begin{cases} \rho_{mn}(3.25 + 0.75\rho_{mn}), & 0 \leq \rho_{mn} \leq 1 \\ \rho_{mn}(3.27 + 0.71\rho_{mn}), & -0.5 \leq \rho_{mn} < 0 \end{cases}$$

where ρ_{mn} denotes the correlation between the null distributions associated with $I[M_m; T_j|T_i]$ and $I[M_n; T_j|T_i]$. P-values obtained from the chi-squared distribution where corrected by FDR using qvality [38], and candidates with estimated q-value below 1E-05 were selected. The procedure produces predicted directional interactions and the set of miRNAs that are predicted to mediate these interactions. Predicted interactions, in at least one direction, are taken as evidence for regulation of the target and regulator genes by their predicted mediators.

In total, evidence for competition for miRNA regulation in TCGA breast cancer tumors supported functional regulation for 299 thousand miRNA-target candidate interactions. Because of the stringent criteria, all candidate functional interactions had interaction scores greater than 0.8, and so, while not by design, all functional interactions where also selected in Phase II; see Figure 2.2(C). In addition to evidence for competition, which is used in Cupid Phase III, interactions with scores greater than 0.5 were for evidence for indirect regulation by miRNAs as described in sections that follow.

2.3 Quality of binding site selection

Cupid evaluates and rescores candidate miRNA binding sites that were predicted by other methods. Here, I used predictions by TargetScan, miRanda and PITA because of source code availability and because their miRNA binding sites prediction methods are complementary; TargetScan is informed by the structure of miRISC, miRanda locally aligns miRNA and target sequences and estimates their binding energy, and PITA uses predicted RNA structure. To evaluate the performance of Cupid binding site scoring and selection, I compared its ability to predict AGO localization in HEK293 [7]. Namely, I tested the ability of

binding site prediction methods to identify 6,905 41-base crosslink-centered regions (CCRs) [7] in 3' UTRs of 3,489 genes, considering both sensitivity and precision of site discovery for the highest expressed miRNAs in HEK293. I tested both the accuracy of binding-site prediction scoring (Figure 2.4(A)) and the effects of miRNA expression (Figure 2.4(B)) on AGO localization for Cupid, TargetScan [23], miRanda [24], PITA [25], DIANA-microT-CDS [42], ElMMo [43], miRmap [44], mirSVR [45], RepTar [46], RNA22 [10], RNAhybrid [47] and TargetSpy [48].

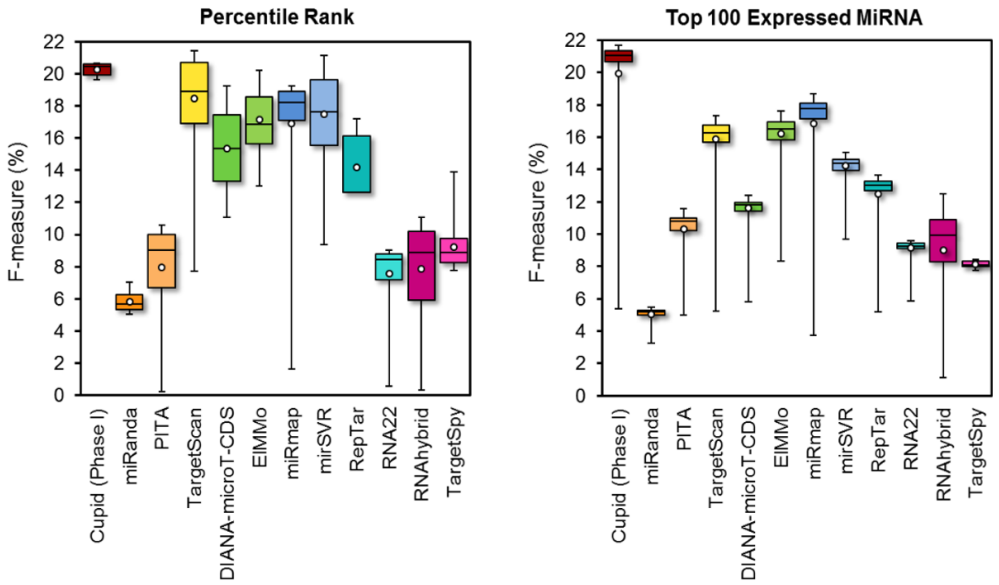


Figure 2.4. Site prediction

(A) Cumulative distributions across ranks of predicted miRNA binding sites that coincide with crosslink-centered regions (CCRs) identified in *EIF2C2* (AGO2) PAR-CLIP experiments, showing average and upper and lower quintiles, and the range of the harmonic mean of precision and recall (F measure) for a panel of target prediction methods. (B) Cumulative distribution across the highest expressed miRNAs of predicted binding sites that coincide with CCRs.

Crosslink-centered regions (CCRs) [7] were taken from Hafner *et. al.* without modification. Predicted sites that overlapped CCRs by at least one base were considered true positive predictions. When multiple miRNAs were queried and predicted binding sites for two miRNAs overlapped the same CCR, both were taken as true positive predictions. Cumulative distributions for score-percentile F measure were generated by first partitioning all predicted sites into 100 equal-size bins, ordered by decreasing confidence, and then calculating the F measure in 100 iterations, where the F measure is evaluated for the top *n* bins at iteration

n. Cumulative distributions for the top 100 expressed miRNAs was generated by calculating the F measure in 100 iterations, where the F measure is evaluated for the *n* highest expressed miRNAs at iteration *n*. The F measure is a harmonic mean of the precision (*P*) and recall (*R*), calculated as

$$F = 2\frac{P \times R}{P + R}$$

Where precision is the fraction of sites that overlap CCRs relative to the total number of predicted sites, and recall is the number CCRs overlapping predicted sites relative to the total number of CCRs.

To study the accuracy of binding-site prediction scoring for each algorithm, I ranked its predicted binding sites for the 100 highest expressed miRNAs in HEK293 according to their score, and constructed cumulative F-measure distributions across these bins, starting from the top scoring sites (Figure 2.4(A)). I compared the F measure across 100 cumulative distributions for each algorithm, plotting, for *k* ranging from 1 to 100, the F measure for predictions in the top *k* ranking bins. The F measure is calculated as the harmonic mean of precision and recall, where precision is the frequency that predicted binding sites overlap CCRs, and recall is the frequency that CCRs overlap predicted binding sites. The F measure was chosen because it eliminates the need to estimate the true negative rate of miRNA binding site prediction and because it balances precision and recall [49]. Given that the number of CCRs is relatively small, I expect the F measure to grow with the number of binding sites, and so methods that predict fewer targets are at a disadvantage. Remarkably, Cupid predictions consistently outperformed predictions by other methods when considering both high- and low-scoring sites and even its most inclusive settings provided a good balance between precision and recall. In comparison, TargetScan and mirSVR performed poorly when only top scoring sites were selected.

To study the effects of miRNA expression on my ability to predict CCRs using miRNA binding sites, I ranked miRNAs according to their expression in HEK293 [7] and constructed cumulative F-measure distributions for CCR-predictive performance using their predicted binding sites, starting from the highest expressed miRNA and up to the 100-highest expressed miRNAs. In Figure 2.4(B) I plot, for *k* ranging from 1 to 100, the F measure for binding-site predictions for the *k*-highest expressed miRNAs. To avoid miRNA-specific scoring biases, I included all predicted sites for each miRNA. Predictions based on a single miRNA are unlikely to explain the majority of AGO binding sites, and I expected to see rapid initial improvement as the number of miRNAs included in the analysis increased, followed by a decline in predictive performance as

low expressed miRNAs are included. Cupid's predictive ability was peaked at an F measure of 21.6% when considering the top 60 expressed miRNAs, followed by miRmap and ElMMo with F measures of 18.6% and 17.3% when considering the top 57 and top 51 miRNAs, respectively. Moreover, the predictive performance of these three algorithms approached its peek once binding site predictions for the 25 highest-expressed miRNAs were included. The results suggest that Cupid-predicted binding sites are in better agreement with AGO binding according to PAR-CLIP data, and that accurate binding site predictions for the top 60 miRNAs are sufficient for identifying AGO binding regions. Including binding site predictions for lower expressed miRNAs in the analysis reduced precision and F measure for all prediction methods.

2.4 Quality of interaction prediction in breast cancer cell lines

To study my ability to predict functional interactions in breast cancer, I used data from three studies that provide genome-level expression fold-change measurements in response to miRNA over expression. The data includes gene expression profiling, in at least two biological replicates, following transfection of pre-mir-18a, pre-mir-193b, pre-mir-206, pre-mir-302c [20], pre-mir-101-1 [21] and scrambled controls in MCF7, and pre-miR-145 and control in MDA-MB-231 [22]. In total, across the six miRNA precursors, I identified 869 down-regulation events of at least 1.4 fold change ($\log_2$ fold change of -0.5), signaling that the corresponding genes may be downstream of transfected miRNAs [50]; see Figure 2.5(A). Considering all binding-site predictions by Cupid (Phase I), I identified 11.3 thousand candidate interactions for miRNAs that may be derived from the six precursors. In total, Cupid predicted 5.7 thousand miRNA-target interactions (Phase II) for these miRNAs, and 3.5 thousand of these had evidence for mRNA competition for miRNA regulation (Phase III). Considering all predicted targets, I calculated the F measure for each method, assuming that predictions including targets that are not down regulated are false positive predictions and that down regulated targets that are not predicted by this method are false negative predictions. Results, given in Figure 2.5(B), suggest that both Cupid's interaction prediction (Phase II) and functional-interaction prediction (Phase III) are significantly better at identifying down regulated genes when compared to Cupid's site prediction (Phase I) ($P<0.01$ according to a contingency table test). Moreover, Cupid's binding site prediction was significantly better at identifying down regulated targets than the next best algorithm (miRmap) at $p<0.05$.

When testing interactions using gene expression data, the F measure for each interaction prediction method was calculated by comparing to genes that were down regulated after transfection of precursors of predicted miRNA regulators. Here, for a transfection experiment with a miRNA precursor, I say that genes with $-\log_2$ expression fold change greater than 0.5 were down regulated. Then, precision was calculated as the fraction of predicted targets of miRNAs derived from the precursor that were down regulated. Similarly, recall was computed as the fraction of down regulated targets that was predicted to be regulated by at least one precursor-derived miRNA.

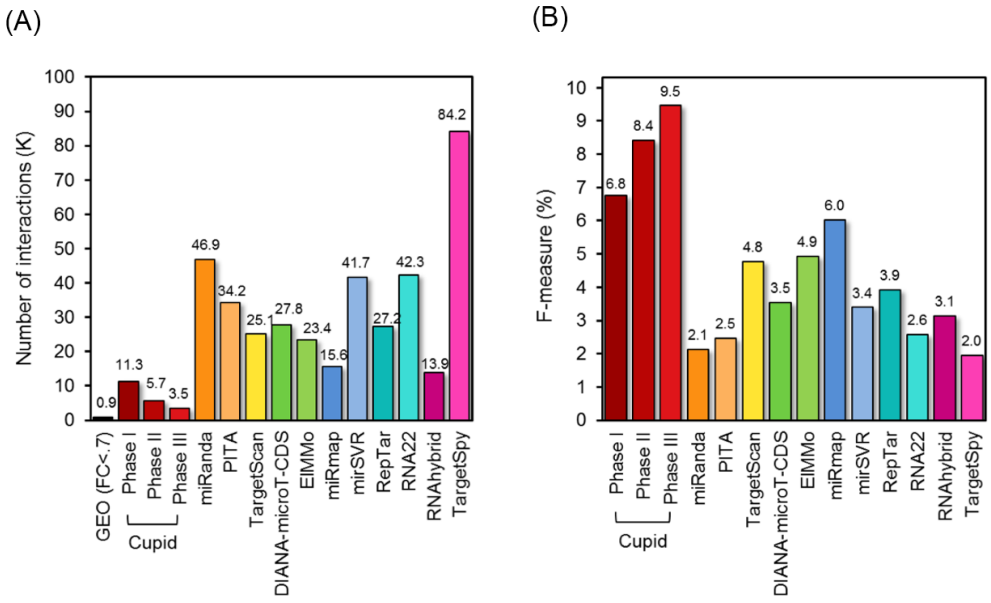


Figure 2.5. Interaction prediction

(A) Number of predicted miRNA-target interactions that were tested through miRNA transfections in MCF7 or MDA-MB-231 for a panel of target prediction methods, including the three phases of Cupid. (B) F measure for the predictive ability of miRNA-target interaction candidates, as predicted by each method, on gene response to miRNA transfections.

2.5 Protein expression tests quality of predictions

To further test predictions, I used RPPAs to measure the response of 120 genes to transfections of 159 miRNA mimics, including 4 mock controls, in MDA-MB-231 using 158 antibodies. Mimics were chosen from a preliminary test of 879 mimics, identifying transfections that lead to highest total fold change across profiling antibodies. Of the 158 RPPA antibodies used, 117 antibodies that correspond to 82 genes were profiled after transfection of at least one of 127 Cupid Phase III-predicted miRNA regulators, with no

replicates. In total, I tested nearly 2,200 interactions predicted by Cupid Phase I, and almost 800 interactions predicted by Cupid Phase III (Figure 2.6(A)).

I first measured the average reduction in protein expression (1 – fold change) relative to mock transfections for predicted interactions by each method, including the three phases of Cupid (Figure 2.6(B)). Cupid Phase III predictions improved on Phase II and Phase I, and were significantly better than the next best prediction method ($p<1E-4$ by Student's t-test compared to DIANA-microT-CDS). Consequently, I focused my detailed analysis on Cupid Phase III predictions.

To test prediction accuracy, I considered each profiling antibody independently, and evaluated protein-expression fold changes after transfection of predicted regulators; results are given in Figure 2.6(C). I plot average fold change, and corresponding Student's t-test derived p values, of protein expression in response to transfection of mimics of predicted miRNA regulators relative to mock transfections. Of the 117 antibodies tested, 34 measured significant ($p<0.05$) fold change reduction in protein expression, and 2 measured significant fold change increase. Of the remaining 81 antibodies, 51 showed reduction and 30 showed increase in protein expression after transfection of predicted regulators, suggesting overall down regulation by predicted miRNAs at $p<4E-10$ by one-sample Kolmogorov-Smirnov test. In total, at a $p<0.05$ by t test, 237 predicted interactions had evidence for reduction in target protein expression after regulator transfection, while 76 showed an increase in target protein expression ($p<6E-84$ by Kolmogorov-Smirnov test). The tests identified ten genes, all previously implicated in regulation of breast cancer tumors, for which RPPA-estimated protein expression reduction averages across mimics of all predicted miRNA regulators were consistently above 10%; see Figure 2.6(D).

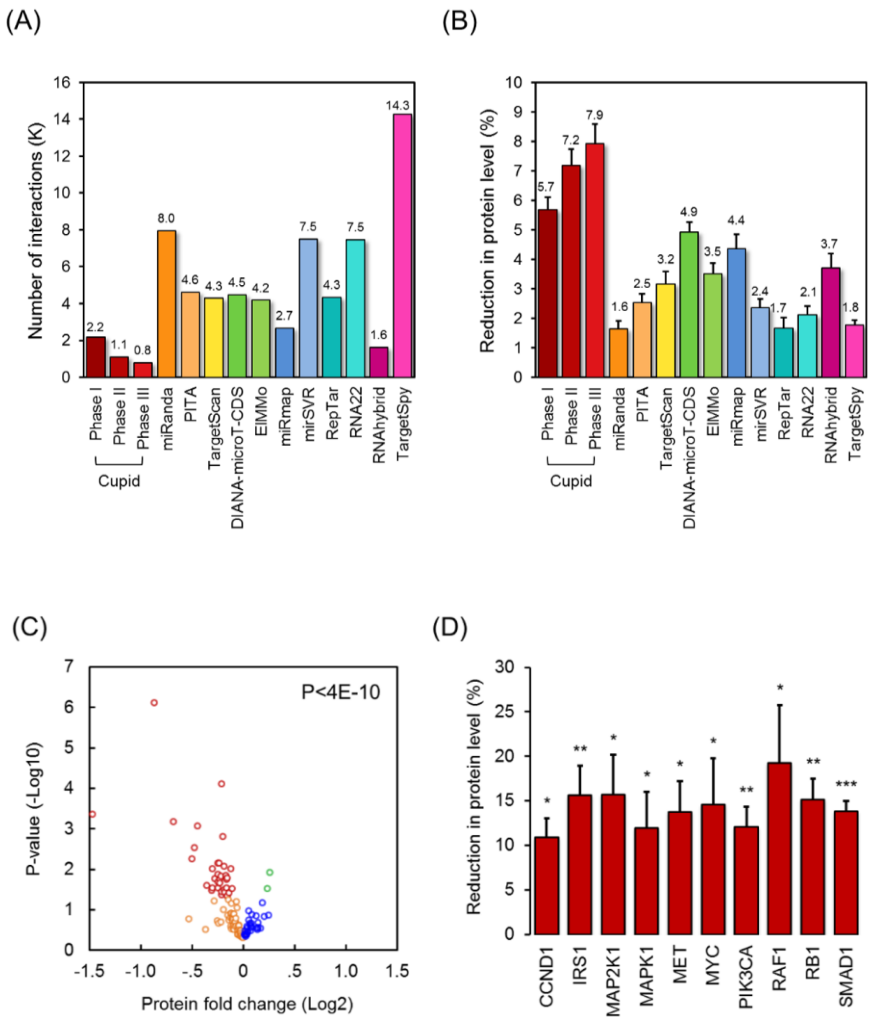


Figure 2.6. High-throughput perturbation tests using protein expression profiling

(A) Number of predicted miRNA-target interactions that were tested through miRNA mimic transfection followed by protein expression profiling. (B) Average reduction in protein level following transfection of the predicted targeting miRNA for a panel of target prediction methods, including the three phases of Cupid. (C) P-values and average protein-expression fold changes after transfection of Cupid-predicted miRNA regulators. In total, considering expression estimates made with 117 antibodies, 34 reported significant down regulation ($p < 0.01$, in red), 51 reported down regulation (orange), and 30 reported up regulation, for a comprehensive significance of $p<2E-08$. (D) Estimated average reduction in protein expression levels for known breast cancer regulators from (C). Data are represented as mean ± SEM.

2.6 Evidence for competition for miRNA regulation

To test predicted functional interactions with evidence for competition for miRNA regulation, I chose to focus on five genes that are known to regulate breast cancer tumors and that were predicted to compete for miRNA regulators (Figure 2.7(A)). To test the ability of the 3' UTRs of CCND1, ESR1, HIF1A and PDGFRA to regulate each of these genes, in addition to NCOA3, my colleagues transfected 3' UTRs and measured gene expression fold changes in MCF7; my colleagues note that they failed to clone the NCOA3 3' UTR and did not test its regulatory potential, and that regulation of PDGFRA was not tested because it was not expressed in the MCF7 cells. Results are given in Figure 2.7(B)-(E), and demonstrate the potential of the 3' UTRs of these genes to regulate mRNA expression in breast cancer. In total, 8 of the 11 predicted directed interactions tested showed significant up regulation in target mRNA expression in response to regulator 3' UTR transfection in MCF7. While this evidence supports regulation by these 3' UTRs [36], it does not identify the miRNAs that these genes compete for.

Hermes predicts that CCND1, ESR1, HIF1A and PDGFRA compete for several common miRNAs, including seven miRNAs that are predicted to target at least three of these four genes. My colleagues used 3' UTR luciferase reporter assays, following mimic transfections, to test that the miRNAs that are predicted to mediate competition by these genes indeed regulate their 3' UTRs. In total, they predicted 30 functional interactions between 10 selected miRNAs and the 3' UTRs of these four genes (Cupid Phase III); predictions and results are given in Figure 2.8. Of particular interest, ESR1, HIF1A and PDGFRA were predicted to compete for hsa-miR-17-5p, miR-106b-5p, hsa-miR-130a/b-3p and hsa-miR-301a-3p. Our assays tested 44 interactions, including interactions with miR-557, which was not predicted to regulate the 3' UTRs. Of the 30 predictions, only regulation of the HIF1A 3' UTR by miR-93-5p was not supported by our assays, suggesting high precision for Cupid's functional-interaction prediction. The remaining assays tested 14 negative predictions, and the results suggest that 8 of the 14 have regulatory potential. In total, our assays suggest that Cupid predictions are accurate, with high precision and good, albeit lower, recall. Precision, if previously validated interactions are included in the analysis, was above 95% while recall was above 75%. Of the 30 predictions, 10 were previously validated, but even after excluding these, Cupid calls were predictive of assay results at $p<0.01$ by Fisher's exact test. Interestingly, one previously reported

interaction, CCND1 regulation by miR-34b, was not predicted by Cupid and was not supported by our luciferase assays.

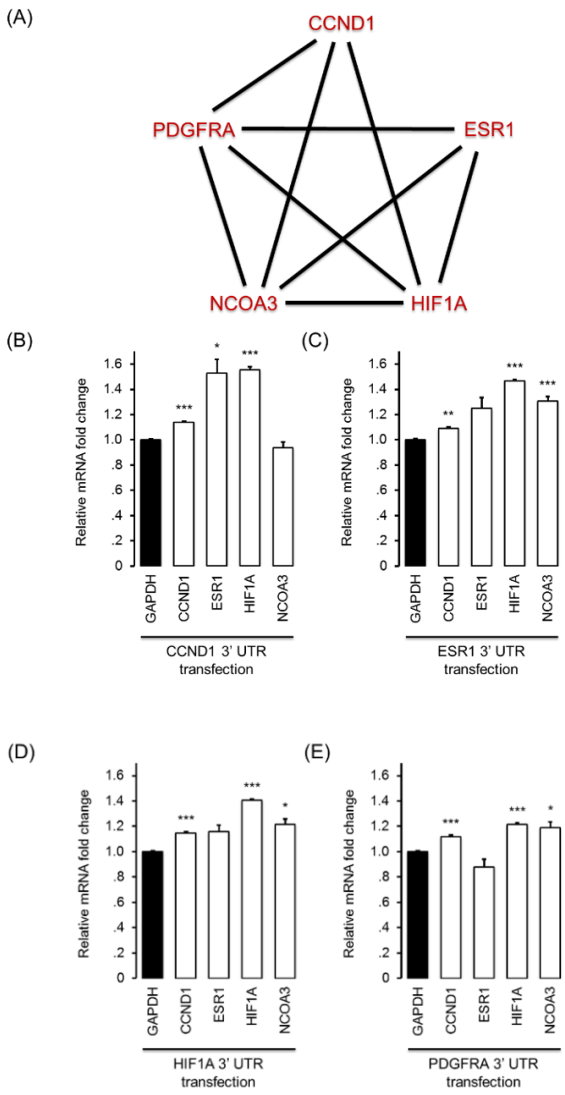


Figure 2.7. Competition for miRNA regulation

Cupid Phase III relies on evidence for competition for miRNA regulation. (A) A subnetwork of oncogenes implicated in breast cancer regulation that were predicted to compete for miRNA regulation with one another. Transfection of the 3' UTRs of (B) CCND1, (C) ESR1 (D) HIF1a, and (E) PDGFRA in MCF7 demonstrates their regulatory potential by up regulating mRNA expression within the subnetwork, as measured by qPCR. Data are represented as mean ± SEM; * p<0.05, ** p<0.01, *** p<0.001.

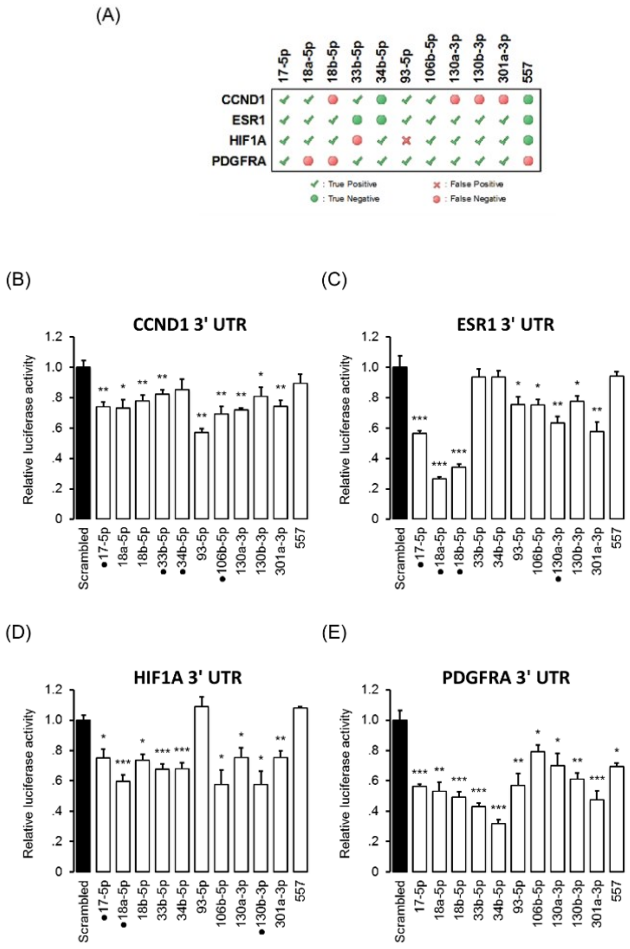


Figure 2.8. Regulatory potential of miRNA modulators

CCND1, ESR1, HIF1A, and PDGFRA were predicted to compete for miRNA regulation, including some miRNAs with evidence for competition by multiple gene pairs. My colleagues tested the regulatory potential of ten of these miRNAs biochemically, with miR-557 selected as negative control. (A) Predicted miRNA-target interactions and a summary of biochemical validation, depicting true positive, true negative, false positive and false negative predictions; down regulation of 3' UTR luciferase activity in response to miRNA-mimic transfection at $p<0.05$ was taken as evidence for regulation. Luciferase activity after miRNA mimic transfection relative to transfection of scrambled control for (B) CCND1, (C) ESR1 (D) HIF1a, and (E) PDGFRa 3' UTRs. Punctuated mimics, for example miR-17-5p for CCND1, ESR1, HIF1a but not PDGFRa, correspond to previously validated interactions. Data are represented as mean ± SEM; * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

2.7 Evidence for indirect regulation for functional regulation

I tested candidate miRNA interactions from Cupid Phase II for additional evidence for indirect regulation by miRNAs. Evidence for indirect regulation by miRNAs examines the correlation between the expression of the miRNA and a set of predicted indirect targets; these were not used to predict direct miRNA-target

interactions and are taken as complementary evidence. In total, this line of evidence produced fewer predictions than the number of predictions with evidence for competition for miRNA regulation. Moreover, my tests based on miRNA perturbations suggest that the predictive ability of this line of evidence is weaker than that of Cupid Phase III. Consequently, I chose to describe them independently. When combined, this lines of evidence support 40,000 predicted interactions that were not selected in Phase III. I first outline the methods and then describe analysis that suggests that this lines of evidence are significant, albeit, weaker classifiers of miRNA regulation.

Evidence for indirect regulation can help identify miRNA targets whose regulation is harder to detect using RNA expression profiles alone. Considering each miRNA m and a predicted direct target T_i^m, I looked for correlation between the breast cancer expression profile of miRNA m and the expression profiles of predicted (direct or indirect) targets of T_i^m; I term T_i^m ***effector***, and its predicted downstream targets are its *regulon*. The total RNA abundance of the regulon may be affected following miRNA-mediated inhibition of the effector, even if the effector's RNA expression is only weakly perturbed. In total, I found evidence for indirect regulation for over 10K predicted interactions. Tested miRNA-target interactions include interactions between miRNAs and target transcription factors with ARACNe-predicted regulons [51], or with genes perturbed by shRNA in Library of Integrated Network-based Cellular Signatures (LINCS) [52]. ARACNe predicted nearly 198,000 interactions with 2,447 transcription factors. The LINCS database includes Luminex-based gene expression fold-change estimates for 1171 genes in response to shRNA-mediated silencing of 1,721 and 1,750 genes at 96 and 144 hours after silencing, respectively, in the luminal breast cancer cell line MCF7. Considering each perturbed effector in LINCS, I attempted to construct a regulon for this effector by selecting profiled genes that responded strongly to its perturbation, relative to both other profiled genes and to responses of these regulon candidates to other perturbations.

To collect evidence for indirect regulation by miRNAs, I first built regulons – sets of predicted direct or indirect targets – for transcription factors expressed in TCGA breast cancer tumors and cancer genes perturbed in LINCS. The expression profiles of these were tested for multivariate correlation, measured by normalized mutual information, with the expression of the miRNA that was predicted to target the transcription factor or cancer gene (effector) upstream from the regulon. Regulons for 2,447 transcription factors were predicted using ARACNe [51], measuring mutual information using adaptive partitioning, with

interaction p-value cutoff 1E-07, DPI coefficient 0, and using consensus predictions from 100 bootstraps. Regulons for genes perturbed by three or more targeting shRNAs in LINCS were collected by identifying genes with high and low fold change in response to shRNA transfection relative to both other profiled genes in response to the same perturbation and to the gene's responses to other perturbations. Significance was measured considering $\log_2$ of the distribution of fold changes, and these were required to be at least 2.5 standard deviations (STD) from mean on either tail of the distribution. However, because the mean fold change μ was not centered at 0, I set the fold change cutoff to $\pm(|\mu| + 2.5 \times \mathrm{STD})$. This produced a more conservative selection that guarded against perturbations and genes with skewed fold change responses.

To measure the correlation between miRNA expression profiles and the expression profiles of its predicted indirect targets, I first computed NMI between the miRNA expression profile in TCGA breast cancer tumors and the expression profiles of all transcribed genes. I compared the vector of NMI values associated with predicted indirect targets (the regulon) to the NMI values for all other genes. The comparison used a running sum statistic based on Fisher's exact test, where I compared, for decreasing NMI cutoffs within the regulon, the number of included and excluded regulon genes and non-target genes. To correct for multiple testing, I used Bonferroni correction for the p-value obtained from the *n*th iteration of the test, considering this p-value as a selection from *n* trials.

For ARACNe-predicted regulons, I used a p<1E-10 significance cutoff, thus correcting for testing miRNA targeting of 2447 regulons, and a total of 2,980,446 tests – up to 2,447 transcription factors targeted by up to 1,218 miRNAs – to obtain an FDR of 1E-03 This correction was not possible for LINCS-derived regulons, which were much 4~5 smaller on average than ARACNe-derived regulons. Thus to select a p-value cutoff I used an optimization procedure based on Bayes' theorem.

Considering D, the set of miRNA-target interactions that have been previously shown, and T, the set of interactions that I have predicted under a specific significance cutoff, then according to Bayes theorem, I can set $P(D|T)P(T) = P(T|D)P(D)$. I seek to maximize precision – the probability that a predicted in interaction is verified, $P(D|T)$. By Bayes theorem, I seek to maximize $P(T|D)/P(T)$. Following this process, I selected a p-value cutoff of 0.01 for both profiling time points, as shown in accompanying figures. The details are depicted in Figure 2.9.

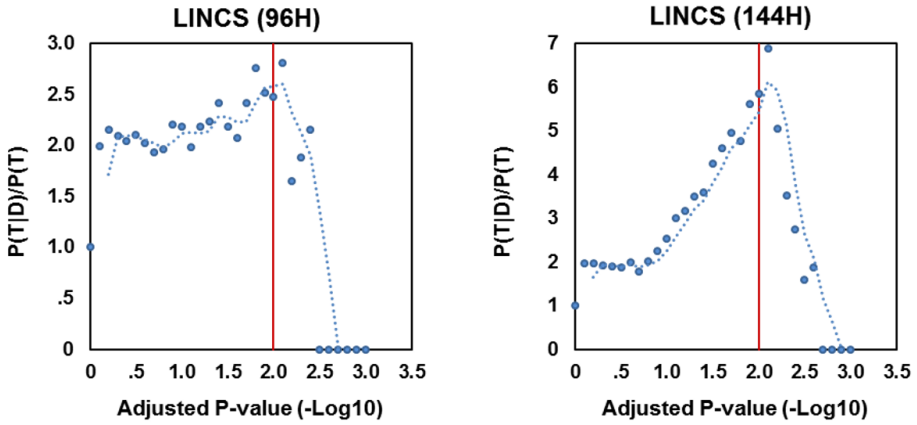


Figure 2.9. Selection of p-value cutoff using Bayes theorem for LINCS data in two time points, 96H and 144H.

To evaluate the significance of the multivariate correlation between the expression profiles of a miRNA and its potential indirect targets, I calculated the enrichment of their normalized mutual information values relative to normalized information values of the expression profiles of the miRNA and all profiled mRNAs. Significantly elevated normalized mutual information values support indirect regulation by the miRNA and provide evidence that it regulates the effector in breast cancer tumors.

I compared the success rate of my method, inferring indirect miRNA regulation based on regulons from ARACNe and LINCS, in identifying true miRNA-target interactions within previously validated interactions, interactions predicted by Cupid, and candidate interactions derived from predictions by TargetScan, miRanda and PITA. For each candidate miRNA-effector interaction, I evaluated the significance of the multivariate correlation between the expression profiles of the miRNA and genes in the effector's regulon. Figure 2.9 depict the frequency of significant correlations between miRNAs and regulons, as a function of significance cutoffs. Results, given in Figure 2.10, suggest that indirect regulation is significantly more likely for true miRNA interactions. In total, expression profiles of miRNAs and regulons, corresponding to miRNA-effector interactions predicted by Cupid, were significantly more likely to be correlated than those predicted using TargetScan, miRanda and PITA sites; $p<4E-30$ and $p<4E-06$, for ARACNe and LINCS regulons, respectively. Previously validated miRNA-effector interactions were the most likely to have significant

miRNA-regulon correlations, but only 151 and 202 miRNA-effector interactions were tested for ARACNe and LINCS regulons, respectively.

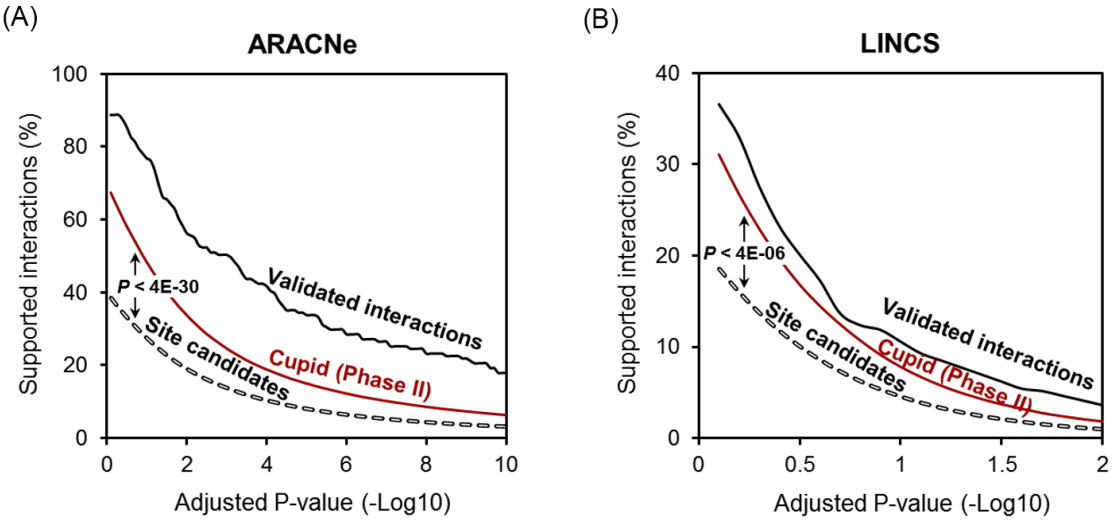


Figure 2.10. Interaction predictions with evidence for indirect regulation

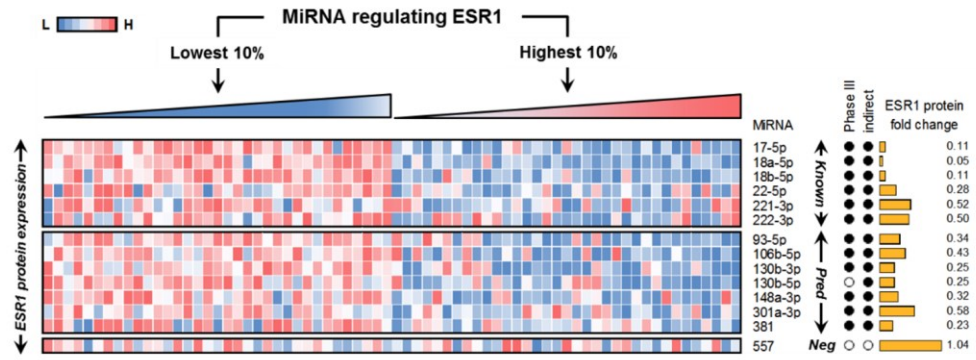
(A) expression-based evidence for regulation of transcription factors through miRNA-expression correlation with the expression of predicted transcription-factor targets in breast cancer tumors, and (B) genes that respond to silencing of predicted miRNA targets in MCF7. I plot the frequency of tested candidate interactions with evidence for indirect regulation by miRNAs as a function of their p-value significance cutoff. Site candidates were drawn by chance from predictions by TargetScan, miRanda and PITA.

I chose to focus on ESR1 for detailed validation. ESR1 regulons were constructed using both ARACNe and LINCS predictions, resulting in evidence for indirect regulation for 44 candidate ESR1 regulators; 8 of these candidate miRNA regulators were shown to regulate ESR1 3' UTR luciferase activity in Fig. 2.8(C). The analysis of ESR1 protein expression in TCGA breast cancer tumors, profiled by RPPAs using the antibody ER.alpha.R.V_GBL.9014870, suggests that ESR1 expression is strongly correlated with the expression of these predicted miRNA regulators, as estimated by miRNA-seq. Biochemical validation of select ESR1 miRNA regulators showed that all but one of the miRNA regulators with evidence for indirect regulation of candidate ESR1 targets significantly reduced ESR1 3' UTR luciferase activity. Details follow below.

To computationally test their potential for functional regulation, I compared ESR1 protein expression in 352 samples with low (bottom 10%) and high (top 10%) expression of each of the fifty candidate miRNA

regulators with evidence for indirect regulation of ESR1 regulons. After removing outliers, for forty four miRNAs within 2 interquartile ranges from the mean, ESR1 protein expression was 2.8 fold higher in samples with low targeting miRNA expression, on average. Results for thirteen selected candidate ESR1 regulators, and miR-557, which was chosen as negative control, are given in Figure 2.11(A) and show significant ESR1 protein expression fold change. Eight of these regulators were selected because their effect on ESR1 3' UTR luciferase activity assays was tested in Figure 2.8(C). The other five regulators where chosen at random from Figure 2.12, and include previously validated regulators miR-22-5p, miR-221-3p and miR-222-3p, as well as previously unknown ESR1 regulators 130b-5p and 148a-3p. Results from biochemical testing of the predicted interactions, including results from assays described in Figure 2.8(C) are given in Figure 2.11(B) for ease of presentation. In total, all but one of the mimic transfections of predicted ESR1 regulators significantly reduced ESR1 3' UTR luciferase activity. While these in vitro assays only measure potential for regulation, my results suggest that over 90% of interactions predicted by Cupid have both regulatory potential and evidence for activity in primary human tumors.

(A)



(B)

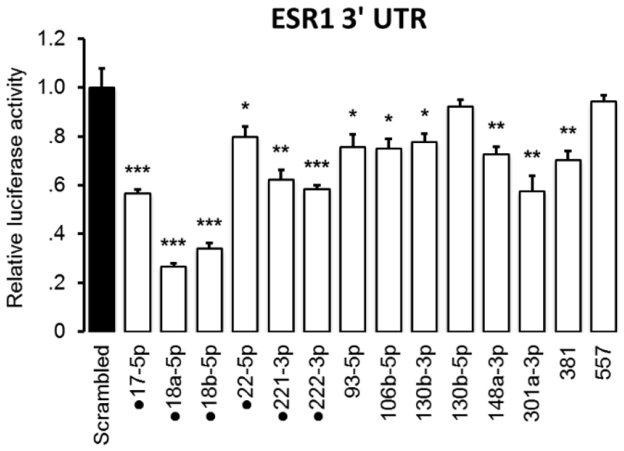


Figure 2.11. Predicted ESR1-regulating miRNAs

(A) *ESR1* protein expression profiles in breast cancer tumors are anti-correlated with the expression of previously validated (Known) and predicted miRNA regulators. Tumors were ranked based on the intensity of the expression of each miRNA, showing *ESR1* relative expression in tumors where candidate regulator expression ranked at the top and at the bottom 10%; each tumor ranking is independent of the other. I mark the presence of evidence for ESR1 competition for each miRNA, evidence for indirect regulation of targets genes downstream from ESR1, and *ESR1* fold change in tumor samples with low versus high expression for each miRNA. (B) Relative 3' UTR luciferase activity in response to mimic transfection. Some data replicated from Figure 2.8. Punctuated mimics correspond to previously validated interactions. Data are represented as mean ± SEM; * p<0.05, ** p<0.01, *** p<0.001.

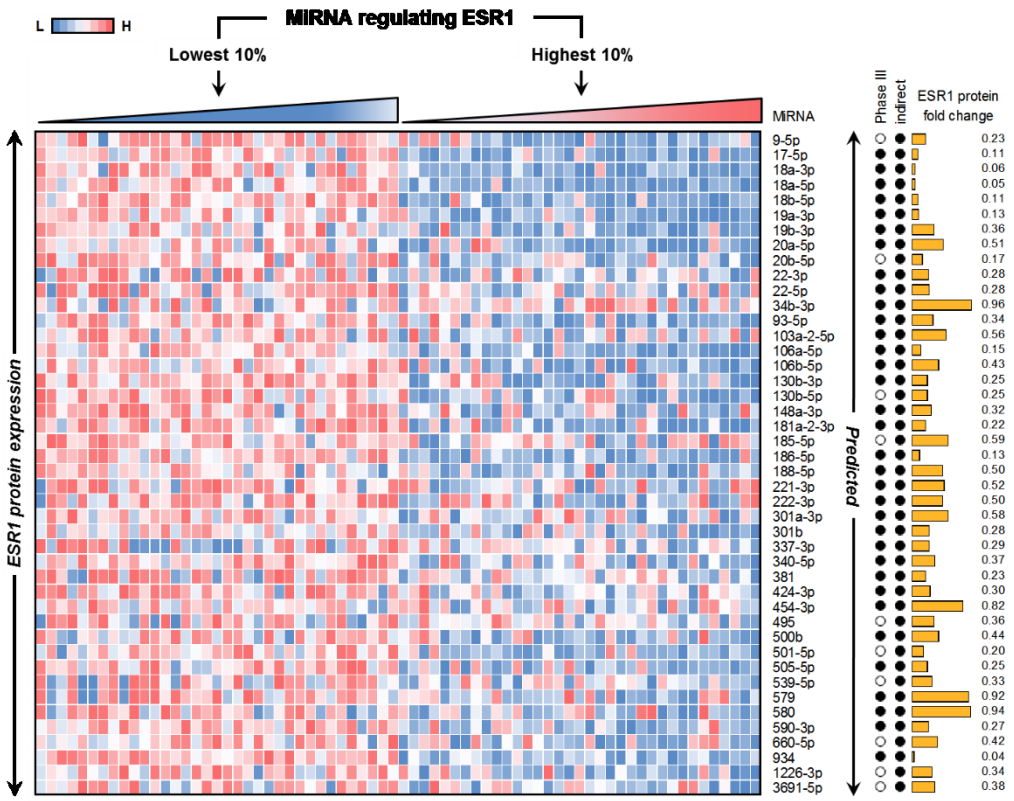


Figure 2.12. ESR1 protein expression profiles support regulation by miRNAs

ESR1 profiles in breast cancer tumors are anti-correlated with the expression of previously validated (Known) and predicted miRNA regulators. I provide data for 44 miRNAs that were predicted to target ESR1 and whose expression was significantly correlated with the expression of ESR1 regulons. The figure follows the design of Figure 2.10(A) and includes miRNA that were not biochemically validated to target ESR1.Tumors were ranked based on the intensity of the expression of each miRNA, showing *ESR1* relative expression in tumors where candidate regulator expression ranked at the top and at the bottom 10%; each tumor ranking is independent of the other. I mark the presence of evidence for ESR1 competition for each miRNA, evidence for synergistic regulation of ESR1, evidence for indirect regulation of targets genes downstream from ESR1, and *ESR1* fold change in tumor samples with low versus high expression for each miRNA.

2.8 Summary

Identifying and understanding pathological implications due to miRNA dysregulation requires accurate maps of functional miRNA targets in specific disease contexts. I describe systems-biology based methods that leverage previously validated interactions together with RNA and protein expression profiles from patient samples to predict functional miRNA-target interactions. Specifically, I focused on predicting functional interactions in breast cancer, using profiles from TCGA breast cancer tumors together with

perturbation data in breast cancer cell lines. A variety of computational and biochemical techniques demonstrated improved fidelity of resulting predictions, including evidence for hundreds of candidate miRNA-target interactions in breast cancer cell lines.

I examined a variety of evidence for functional regulation by miRNAs in breast cancer, including evidence that putative miRNA targets compete for regulation by their common targeting miRNAs, evidence for synergy between miRNA species, and evidence for indirect regulation by miRNAs derived from expression-based correlation between the miRNA and the putative targets of its predicted targets (effectors). I found evidence for competition for nearly 300,000 interactions, and other lines of evidence only marginally added to this set. Moreover, while evidence for synergistic regulation by miRNAs and indirect regulation of effector targets were significantly predictive of true miRNA interactions, they failed to significantly improve the prediction accuracy, as measured through perturbation experiments in breast cancer cell lines. I believe that this evidence will be useful for building predictor functions, but chose not to explicitly include this evidence in my miRNA-target prediction algorithm Cupid at this time.

High throughput data from reverse Phase protein arrays (RPPAs) in disease tissues has been recently made publically available. Focusing on predicted regulators of ESR1, I show that RPPA data in breast cancer tumors could be used as an effective filter for identifying functional miRNA regulators. To further test the effects of miRNA regulation on a select set of proteins, I profiled protein expression after miRNA perturbation, producing a dataset that could be used to compare prediction performance, identifying breast cancer genes that are particularly amendable to miRNA regulation and validating 237 miRNA-target interactions for known cancer genes with evidence for regulation in breast cancer tumors.

Success in efforts to elucidate regulatory interactions that channel the effects of genomic alterations to dysregulate cancer genes requires accurate disease-specific wiring diagrams, including functional miRNA regulatory interactions. Currently, the elucidation of these can only be done using computational approaches, as technical limitations make direct detection of miRNA target impractical. I showed that computational approaches that collect evidence for functional regulation by miRNAs in the given context have significantly improved balance between precision and recall. I showed that evidence for competition for miRNA regulation helps significantly improve miRNA target prediction, and gave evidence that systems-biology approaches may help improve prediction as well. Improving context-specific interaction prediction

will have considerable implications for personalized medicine, and given the increasing body of molecular profiles in primary disease tissues and in perturbation of disease models is an opportunity for systems-biology based approaches to improve accuracy.

References

1. Filipowicz W, Bhattacharyya SN, Sonenberg N: Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nature reviews Genetics* 2008, 9(2):102-114.

2. Garzon R, Calin GA, Croce CM: MicroRNAs in Cancer. *Annu Rev Med* 2009, 60:167-179.

3. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA *et al*: MicroRNA expression profiles classify human cancers. *Nature* 2005, 435(7043):834-838.

4. Mukherji S, Ebert MS, Zheng GX, Tsang JS, Sharp PA, van Oudenaarden A: MicroRNAs can generate thresholds in target gene expression. *Nat Genet* 2011, 43(9):854-859.

5. Liu J, Valencia-Sanchez MA, Hannon GJ, Parker R: MicroRNA-dependent localization of targeted mRNAs to mammalian P-bodies. *Nature cell biology* 2005, 7(7):719-723.

6. Chi SW, Zang JB, Mele A, Darnell RB: Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* 2009, 460(7254):479-486.

7. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M, Jr., Jungkamp AC, Munschauer M *et al*: Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* 2010, 141(1):129-141.

8. Helwak A, Kudla G, Dudnakova T, Tollervey D: Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* 2013, 153(3):654-665.

9. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS: MicroRNA targets in Drosophila. *Genome Biol* 2003, 5(1):R1.

10. Miranda KC, Huynh T, Tay Y, Ang YS, Tam WL, Thomson AM, Lim B, Rigoutsos I: A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* 2006, 126(6):1203-1217.

11. Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB: Prediction of mammalian microRNA targets. *Cell* 2003, 115(7):787-798.

12. Friedman RC, Farh KK, Burge CB, Bartel DP: Most mammalian mRNAs are conserved targets of microRNAs. *Genome research* 2009, 19(1):92-105.

13. Gennarino VA, Sardiello M, Avellino R, Meola N, Maselli V, Anand S, Cutillo L, Ballabio A, Banfi S: MicroRNA target prediction by expression analysis of host genes. *Genome research* 2009, 19(3):481-490.

14. Plaisier CL, Pan M, Baliga NS: A miRNA-regulatory network explains how dysregulated miRNAs perturb oncogenic processes across diverse cancers. *Genome research* 2012, 22(11):2302-2314.

15. van Iterson M, Bervoets S, de Meijer EJ, Buermans HP, t Hoen PA, Menezes RX, Boer JM: Integrated analysis of microRNA and mRNA expression: adding biological significance to microRNA target predictions. *Nucleic acids research* 2013.

16. Selbach M, Schwanhausser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N: Widespread changes in protein synthesis induced by microRNAs. *Nature* 2008, 455(7209):58-63.

17. Basso K, Sumazin P, Morozov P, Schneider C, Maute RL, Kitagawa Y, Mandelbaum J, Haddad J, Jr., Chen CZ, Califano A *et al*: Identification of the human mature B cell miRNome. *Immunity* 2009, 30(5):744-752.

18. Xu J, Li CX, Li YS, Lv JY, Ma Y, Shao TT, Xu LD, Wang YY, Du L, Zhang YP *et al*: MiRNA-miRNA synergistic network: construction via co-regulating functional modules and disease miRNA topological features. *Nucleic acids research* 2011, 39(3):825-836.

19. Boissonneault V, Plante I, Rivest S, Provost P: MicroRNA-298 and microRNA-328 regulate expression of mouse beta-amyloid precursor protein-converting enzyme 1. *The Journal of biological chemistry* 2009, 284(4):1971-1981.

20. Leivonen SK, Makela R, Ostling P, Kohonen P, Haapa-Paananen S, Kleivi K, Enerly E, Aakula A, Hellstrom K, Sahlberg N *et al*: Protein lysate microarray analysis to identify microRNAs regulating estrogen receptor signaling in breast cancer cell lines. *Oncogene* 2009, 28(44):3926-3936.

21. Frankel LB, Wen J, Lees M, Hoyer-Hansen M, Farkas T, Krogh A, Jaattela M, Lund AH: microRNA-101 is a potent inhibitor of autophagy. *EMBO J* 2011, 30(22):4628-4641.

22. Gotte M, Mohr C, Koo CY, Stock C, Vaske AK, Viola M, Ibrahim SA, Peddibhotla S, Teng YH, Low JY *et al*: miR-145-dependent targeting of junctional adhesion molecule A and modulation of fascin expression are associated with reduced breast cancer cell motility and invasiveness. *Oncogene* 2010, 29(50):6569-6580.

23. Lewis BP, Burge CB, Bartel DP: Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 2005, 120(1):15-20.

24. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS: Human MicroRNA targets. *PLoS Biol* 2004, 2(11):e363.

25. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E: The role of site accessibility in microRNA target recognition. *Nat Genet* 2007, 39(10):1278-1284.

26. Cancer Genome Atlas N: Comprehensive molecular portraits of human breast tumours. *Nature* 2012, 490(7418):61-70.

27. Papadopoulos GL, Reczko M, Simossis VA, Sethupathy P, Hatzigeorgiou AG: The database of experimentally supported targets: a functional update of TarBase. *Nucleic acids research* 2009, 37(Database issue):D155-158.

28. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K *et al*: TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic acids research* 2006, 34(Database issue):D108-110.

29. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T: miRecords: an integrated resource for microRNA-target interactions. *Nucleic acids research* 2009, 37(Database issue):D105-110.

30. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ: miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic acids research* 2006, 34(Database issue):D140-144.

31. Chang C-C, Lin C-J: LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2011, 2(3):27:21--27:27.

32. Wu T-f, Lin C-J, Weng RC: Probability Estimates for Multi-class Classification by Pairwise Coupling. *Journal of Machine Learning Research* 2003, 5:975--1005.

33. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S *et al*: Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research* 2005, 15(8):1034-1050.

34. Darbellay G, Vajda I: Estimation of the Information by an Adaptive Partitioning of the Observation Space. *IEEE Trans on Information Theory* 1999, 45:1315--1321.

35. Press W, Teukolsky SV, WT, Flannery B: Numerical Recipes: The Art of Scientific Computing, 3rd edn. New York: Cambridge University Press; 2007.

36. Sumazin P, Yang X, Chiu HS, Chung WJ, Iyer A, Llobet-Navas D, Rajbhandari P, Bansal M, Guarnieri P, Silva J *et al*: An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell* 2011, 147(2):370-381.

37. Alexa A, Rahnenfuhrer J, Lengauer T: Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* 2006, 22(13):1600-1607.

38. Kall L, Storey JD, Noble WS: Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. *Bioinformatics* 2008, 24(16):i42-48.

39. Satterthwaite FE: An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin* 1946 2:110–114.

40. Hou C-D: A simple approximation for the distribution of the weighted combination of non-independent or independent probabilities. *Statistics & Probability Letters* 2005, 73(2):179–187.

41. Brown MB: A method for combining non-independent, one-sided tests of significance. *Biometrics* 1975, 31:987–992.

42. Reczko M, Maragkakis M, Alexiou P, Grosse I, Hatzigeorgiou AG: Functional microRNA targets in protein coding sequences. *Bioinformatics* 2012, 28(6):771-776.

43. Gaidatzis D, van Nimwegen E, Hausser J, Zavolan M: Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC bioinformatics* 2007, 8:69.

44. Vejnar CE, Zdobnov EM: MiRmap: comprehensive prediction of microRNA target repression strength. *Nucleic acids research* 2012, 40(22):11673-11683.

45. Betel D, Koppal A, Agius P, Sander C, Leslie C: Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol* 2010, 11(8):R90.

46. Elefant N, Altuvia Y, Margalit H: A wide repertoire of miRNA binding sites: prediction and functional implications. *Bioinformatics* 2011, 27(22):3093-3101.

47. Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R: Fast and effective prediction of microRNA/target duplexes. *RNA* 2004, 10(10):1507-1517.

48. Sturm M, Hackenberg M, Langenberger D, Frishman D: TargetSpy: a supervised machine learning approach for microRNA target prediction. *BMC bioinformatics* 2010, 11:292.

49. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A *et al*: A large-scale evaluation of computational protein function prediction. *Nature methods* 2013, 10(3):221-227.

50. Guo H, Ingolia NT, Weissman JS, Bartel DP: Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* 2010, 466(7308):835-840.

51. Margolin A, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera R, Califano A: ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC bioinformatics* 2006, 7(Suppl 1):S7.

52. Peck D, Crawford ED, Ross KN, Stegmaier K, Golub TR, Lamb J: A method for high-throughput gene expression signature analysis. *Genome Biol* 2006, 7(7):R61.

Chapter 3: Hermes – reverse engineering ceRNA network in glioblastoma

3.1 Preface

Note that this chapter had been published by Cell Press in October 14th, 2011. The paper title is "*An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma*". According to the editorial policies defined by Cell Press, any author of this paper is allowed to "*include the article in full or in part in a thesis or dissertation (provided that this is not to be published commercially)*"; see the section of Authors' Rights at http://www.cell.com/authors for details. As one of the co-1st authors, I am including this published paper with slight modifications as a part of my dissertation.

3.2 Summary

By analyzing gene expression data in glioblastoma in combination with matched microRNA profiles, we have uncovered a post-transcriptional regulation layer of surprising magnitude, comprising over 248,000 microRNA (miR)-mediated interactions. These include ~7,000 genes whose transcripts act as miR 'sponges' and 148 genes that act through alternative, non-sponge interactions. Biochemical analyses in cell lines confirmed that this network regulates established drivers of tumor initiation and subtype, including *PTEN*, *PDGFRA*, *RB1*, *VEGFA*, *STAT3*, and *RUNX1*, suggesting that these interactions mediate crosstalk between canonical oncogenic pathways. RNA silencing of 13 microRNA-mediated *PTEN* regulators, whose locus deletions are predictive of *PTEN* expression variability, was sufficient to downregulate *PTEN* in a 3' UTR-dependent manner and to increase tumor-cell growth rates. Thus, this miR-mediated network provides a mechanistic, experimentally validated rationale for the loss of *PTEN* expression in a large number of glioma samples with an intact *PTEN* locus.

3.3 Introduction

Dysregulation of physiologic microRNA (miR) activity has been shown to play an important role in tumor initiation and progression, including gliomagenesis [1-5]. Therefore, molecular species that can regulate miR activity on their target RNAs, without affecting the expression of relevant mature miRs, may play equally relevant roles in cancer. Yet, few such modulators of miR-activity have been characterized [6,7], and both

the extent and relevance of their role in controlling normal cell physiology and pathogenesis are poorly understood.

By analyzing a large set of sample-matched gene and miR expression profiles from *The Cancer Genome Atlas (*TCGA), we show here that the regulation of target genes by modulators of miR activity is surprisingly extensive in human glioma and that it affects genes with an established role in gliomagenesis and tumor subtype implementation. We defined s*ponge modulators* (Figure 3.1) that include both messenger RNAs (mRNAs) and noncoding RNAs, which share miR-binding sites with other RNAs targeted by the miR. Thus, these modulators act as miR *sponges* or competitive endogenous RNA (ceRNA) via an established titration mechanism [7-9]. Depending on their expression levels and on the total number of functional miR binding sites they share with a target, sponge modulators can decrease the number of free miR molecules available to repress other functional targets. Established sponge-modulators include *VCAN* [10], *PTENP1* [7] and *CD44* [11].

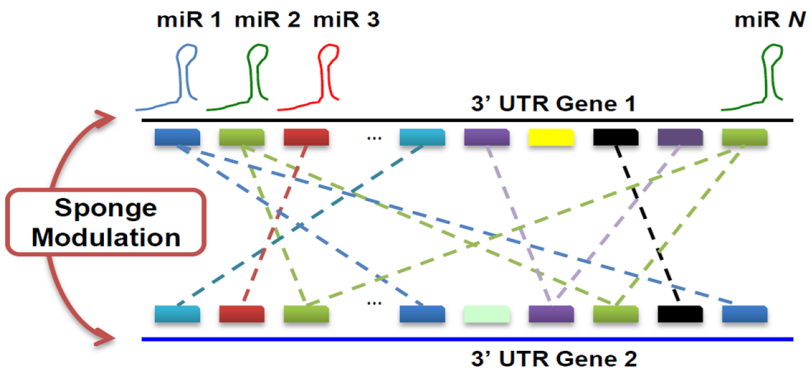


Figure 3.1. Sponge modulators

MiR activity modulation may be implemented by several distinct mechanisms. We consider competition by RNAs for a common miR program (sponge effect) separately from other mechanisms, such as those driven by protein-protein or protein-miR interactions. RNAs modulate each other through their common miR-regulatory program. Up/down changes to the expression of one RNA perturb the relative abundance of functioning miRs that target both RNAs, leading to a corresponding up/down regulation of the second RNA.

To evaluate both the range and potential tumorigenic role of this class of miR-mediated interactions, we present a new multivariate analysis method, called Hermes, which systematically infers candidate modulators of miR activity from large collections of genome-wide expression profiles of both genes and miRs from the same tumor samples. Hermes extends the functionality of MINDy (modulator inference by

network dynamics) algorithm, which uses measurements from information theory to identify genes that modulate transcription factor activity via posttranslational modifications. MINDy has been used to infer post-translational modulators of the MYC transcription factor in human B cells [12], to infer signaling modulators of all transcription factors in human B cells [13], and to identify the ubiquitin conjugating ligase HUWE1 as a modulator of N-MYC turnover in neural stem cells [14].

In essence, MINDy and Hermes make inferences by estimating two quantities from information theory: the Mutual Information (MI) and Conditional Mutual Information (CMI). The MI quantifies how much one variable informs about another variable (i.e., high MI between two variables implies that knowledge about the first variable is predictive of state of the second variable). The CMI calculates the expected value of MI of two variables given the third variable. Specifically, given a modulator (M), a regulator (R), and a regulated target (T), the algorithms dissect the regulatory dependency of these three components by studying the difference between the Conditional Mutual Information (CMI) of the regulator's expression level and the target's expression level, (conditional on the expression level of the modulator) and the Mutual Information (MI) of the regulator and target expressions, $\Delta I = I[R;T|M] - I[R;T]$ [15]. These quantities and their associated statistical significance can be computed from large collections of gene expression profiles (>250 samples), using a variety of estimators for MI and CMI [15], i.e. computational tools that can quantitatively estimate their values.

Hermes expands the MINDy information theoretic framework to identify candidate genes that modulate miR activity (i.e., modulators), whose availability M affects the relationship between the expression of miRs targeting a gene T and its expression profile, T. We use the term *miR program* to indicate a set of miRs targeting a gene and the term *common miR program* to indicate the intersection between the miR programs of two distinct genes. Analysis of Hermes-inferred sponge interactions in TCGA glioblastoma data revealed a regulatory network of previously unsuspected size. Experimental validation of 26 such interactions, of which only 3 failed to validate, suggested that Hermes has a low false positive rate and showed that mPR interactions participate collectively in regulation of key drivers of gliomagenesis and tumor subtype, that these interactions mediate cross-talk between independent pathways, and that they affect cell pathophysiology.

3.4 Hermes framework

While MINDy considers one candidate modulator/regulator/target triplet at a time, Hermes integrates the analysis across all miRs in the common miR program of two genes, using Fisher's method [16]. Specific technical details of the analysis are provided in Experimental Procedures. The cartoon example of Figure 3.2 illustrates the type of interaction that Hermes can help dissect. Here, the increase in expression of the modulator gene is associated with a corresponding increase in mutual information between the expression of several miRs and the expression of their common target.

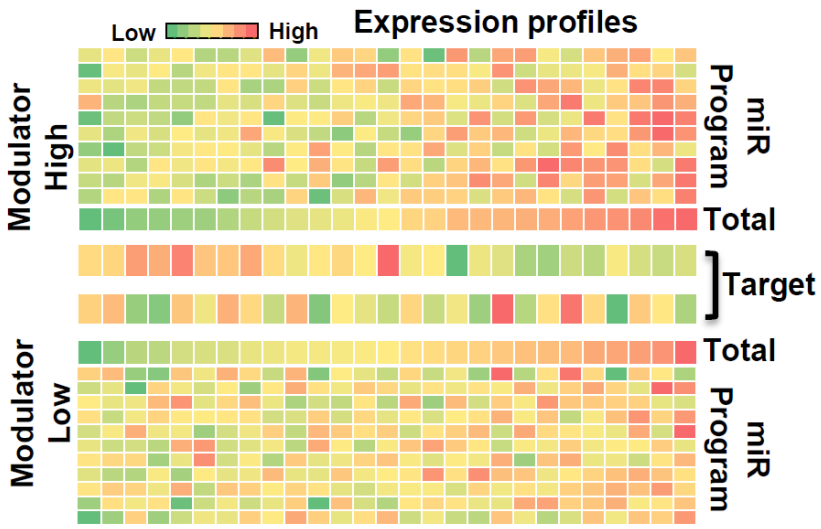


Figure 3.2. Identification of sponge modulators using conditional mutual information (CMI)

To identify candidate modulators, we sought out instances where the correlation between the total expression of a miR program and its target is dependent on the expression of a candidate modulator. This image visualizes a simplification of the process. The top heatmap shows expression of miRs in a program (rows) across all samples (columns) where the modulator expression is high, with the bottom line showing the total expression of the miR-program in the sample. Samples are sorted low to high based on miR-program expression. Below that is the expression of the target of the miR-program. The top heatmap shows strong inverse correlation between miR-program expression and target expression, consistent with an active miR program. The bottom heatmap shows the same data but this time for samples where modulator expression is low. Here, the negative correlation between miR-program expression and target expression is reduced, which is indicative of a suppressed miR program.

In principle, one could evaluate all possible modulator/miR/target triplets and then select statistically significant ones that share the same modulator and target via different miRs. While this would avoid having to select relevant miR programs *a priori*, it would also entail evaluating a huge number of triplets (~4.0E+11), which is computationally prohibitive and will effectively prevent the discovery of many interactions due to

excessive multiple hypothesis testing correction. Similar to MINDy, which addresses this problem by testing only triplets for experimentally validated or computationally inferred transcription factor-target interactions, we use a new miR-target discovery algorithm, Cupid, that is specifically tailored to the identification of miR programs to reduce the number of statistical tests performed by Hermes, see Experimental Procedures. Specifically, Hermes considers only modulator-target pairs, (M, T) sharing a statistically significant number of miRs in their Cupid-inferred common miR programs. In addition, Hermes assumes that sponge interactions are symmetric and thus jointly evaluates the statistical significance of both M as a miR-program mediated modulator of T and of T as a miR-program mediated regulator of M, by combining *p*-values using Fisher's method [16]. Indeed, even though miR-binding and regulatory kinetics may differ in the two targets, most sponge-mediated interactions should still exhibit symmetric behavior. This is because, when averaged over the multiple miRs in their common miR program, the differences in the number of individual miR binding sites and their regulatory kinetics should average out. As shown in Experimental Procedures, symmetry analysis confirmed that only a very small fraction of candidate sponge interactions with strictly asymmetric supporting evidence is missed by Hermes (< 0.02%).

3.5 The mPR network

The statistical significance of $\Delta I = I[miR; T|\, M] - I[miR; T]$ can be effectively estimated from a large number of samples (>250), using a variety of CMI estimators [12], provided that matched miR and gene expression profiles are available for the same samples. The Cancer Genome Atlas (TCGA) datasets are thus ideally suited for this analysis as they are among only a handful satisfying these requirements. For this analysis, we used a publicly available set of 262 matched gene (both mRNA and non-coding RNA) and miR expression profiles from glioblastoma biopsies [17]. When used in genome-wide fashion on this dataset, Hermes identified nearly 7,000 sponge modulators participating in ~248,000 pairwise miR-program-mediated (RNA-RNA) interactions at a highly conservative False Discovery Rate (FDR < 1*e*-04). These interactions are summarized in Figure 3.3(A). Sponge interactions constitute a large and previously uncharacterized miR-Program mediated Regulatory (*mPR*) Network.

Globally, the sponge-mediated component of the mPR network presents roughly the same size and scale-free structure of typical transcriptional regulatory networks [18]. For instance, ARACNe-based reverse

engineering of transcriptional interactions in glioblastoma dissected ~150,000 distinct TF-target interactions [19], compared to ~248,000 mPR interactions inferred by Hermes. We modeled the network graphically, with RNAs represented as nodes and their sponge-mediated mPR interactions as undirected edges (Figure 3.3(A)). Since inferred sponge-interactions are symmetric, RNAs in this network both regulate and are regulated by their neighbor RNAs. However, mPR interactions sharing a common RNA do not necessarily interact through the same miR program. Common miR programs supporting mPR network interactions include 18 miRs on average and up to a maximum of 153 miRs. This suggests that, on average, sponge modulation effects associated with each individual miR in a common program may be negligible compared to the global effect of the entire program.

The mPR network contains many highly interconnected (i.e., *dense*) structures, *i.e.*, *N*-gene sub-graphs, with a number of internal edges approximating the theoretical maximum $N_{\max} = N(N-1)/2$. Indeed, the largest dense glioma mPR structure is a 564-node, 111-core sub-graph [20], i.e., a structure where each RNA is directly linked to at least 111 of the other 563 RNAs. RNAs in these dense sub-graphs are strongly co-expressed, since each RNA tracks the average expression of the other sub-graph members it is connected to. Densest sub-graph RNAs and their interactions are shown in red, near the center of Figure 3.3(A). Conversely, sub-graphs near the edge of the figure, which are shown in purple, are sparser and their members are less co-expressed. Nodes in the network are clustered according to their sub-graph connectivity and each node is depicted with a color representing the size of the sub-graph that contains it. The mPR network is scale free, see Experimental Procedures. Thus, the number of same color nodes in a band increases exponentially with their distance from the center (Figure 3.3(A)).

The overall regulatory effect on a node depends on many variables, including the number of its mPR neighbors, the size of the miR programs that mediate its interactions, and the individual kinetics of the individual miR-target interactions it shares with its neighbors. In general, however, nodes in larger highly connected sub-graphs will have more neighbors and will thus be more strongly regulated by their mPR interactions. Indeed, co-expression of RNAs in a sub-graph increases linearly with the sub-graph size, as shown in Figure 3.3(B).

Analysis of the mPR network shows that mPR interactions participate in distal regulation between genes within and across chromosomes, see Figure 3.3(C) by CIRCOS [21]. In addition, analysis of KEGG pathway

genes [22] targeted by mPR interactions shows that this interaction layer mediates cross-talk between numerous pathways.

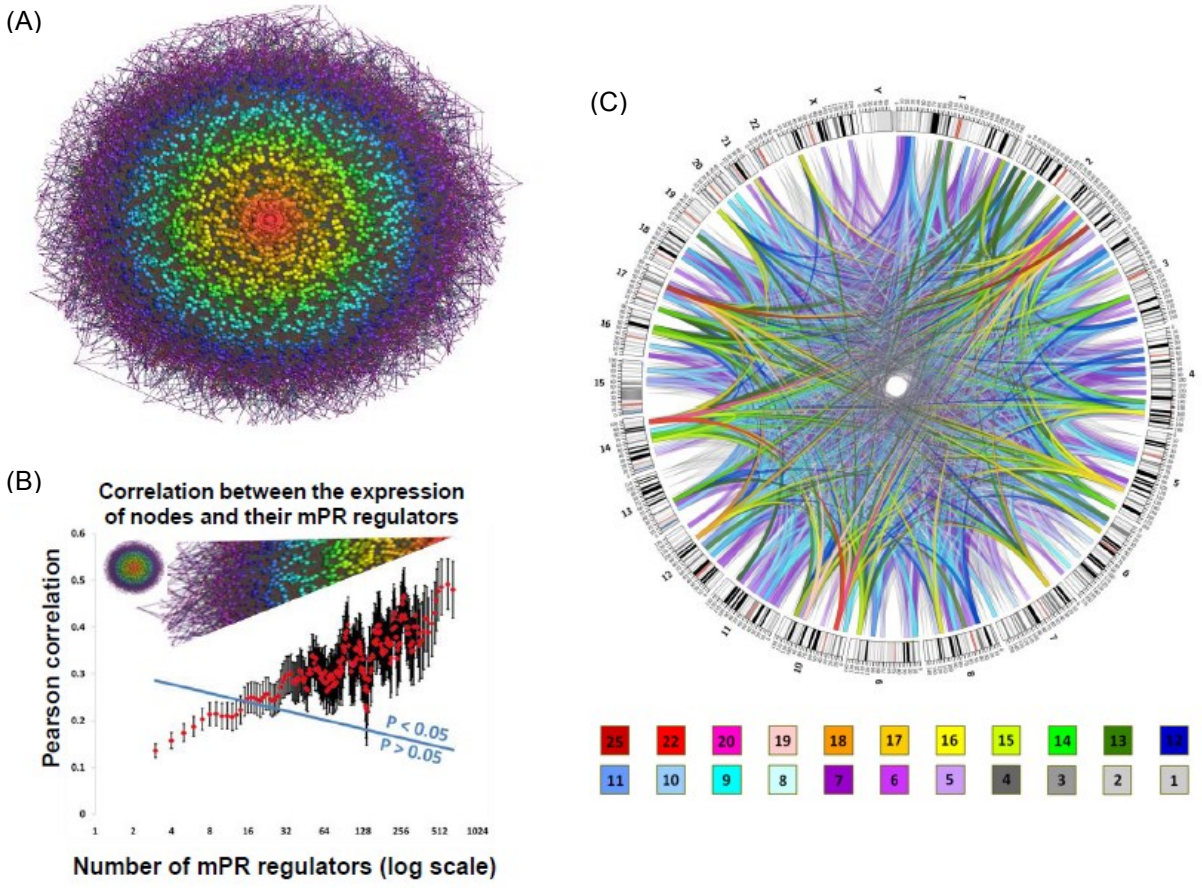


Figure 3.3. The mPR network

(A) Genome-wide inference of sponge modulators identified a miR-program mediated post-transcriptional regulatory (mPR) network including ~248,000 interactions. Its graphic visualization uses nodes to represent individual RNAs and edges to represent miR-program mediated RNA-RNA interactions. Nodes near the center of the graph are contained within more tightly regulated, dense sub-graphs, with the densest 564-node sub-graph shown in red at the center of the network. The network is scale free, and the color bands, which include nodes with similar connectivity, have a size that increases exponentially with the distance from the center. (B) The correlation between expression of RNAs and the total expression of their mPR regulators (i.e., all its mPR neighbors) is plotted as a function of the number of its mPR regulators; genes at the center of the mPR network are regulated by hundreds of mPR regulators and are significantly correlated with their total expression. Values above the blue line are statistically significant at $p < 0.05$. (C) The 564-node mPR sub-graph facilitates interactions between the loci of distal genes. Colors designate the number of gene-to-gene edges connecting each 10Mb chromosomal region.

3.6 PTEN expression is regulated by mPR interactions

PTEN downregulation is a hallmark of gliomagenesis and its locus has been identified as one of the most frequently altered in glioblastoma [23]. While homozygous deletions at the *PTEN* locus are rare, appearing in less than 2% of glioblastoma samples, *PTEN* is haploinsufficient and even moderate *PTEN* downregulation at the protein level, such as that resulting from loss of a single allele, may be tumorigenic. Surprisingly, however, the range of *PTEN* expression in heterozygously deleted samples is comparable to its range of expression in samples where its locus is intact (Figure 3.4 (A), (B)), suggesting that its expression may be tightly regulated and that a variety of additional mechanisms may contribute to its downregulation in tumors. PTEN regulation by miRs is well established. In glioblastoma, for instance, amplifications at the miR-26a locus have been implicated with downregulation of PTEN [3]. Interestingly, *PTEN* is one of the genes in the densest 111-core sub-graph, with a total of 534 interactions in the mPR network, suggesting that its expression is strongly regulated by sponge effect. Indeed, not only do >80% of the tumors with an intact *PTEN* locus have deletions in at least one of the PTEN mPR regulator loci (44/53 as of March, 2011), but the total number of such deletions in each sample is highly predictive of *PTEN* expression (p < 1e-04 by permutation testing), see Figure 3.4 (B). Interestingly, these deletions are more predictive of *PTEN* expression in *PTEN* wild type tumors than in tumors with *PTEN* heterozygous deletions, suggesting that mPR regulators may account for missing *PTEN* genetic variability in glioblastoma.

We focused on a subset of 13 *PTEN* mPR regulators that are expressed in the glioblastoma cell line SNB19 and whose loci are enriched for deletions in tumors with an intact *PTEN*-locus (25/53); interestingly, these 25 tumors include all of the 8/53 *PTEN*-intact tumors with amplification at the miR-26a. The total number of deletions at these 13 gene loci, as well as their total mRNA expression, were found to be highly predictive of *PTEN* expression not only in *PTEN* intact samples but across all 262 tumors tested in our analysis (Figure 3.4(A),(B); p_{Del} < 2e-10 and p_E < 5e-23 by Pearson correlation analysis, respectively). Correlation between the genetics and genomics of *PTEN* mPR regulators and *PTEN*'s mRNA expression suggests that deletions at *PTEN* mPR loci may collectively represent a key contribution to loss of *PTEN* expression in glioblastoma. In addition, *PTEN* is not the only gene whose expression may be regulated by deletions at the loci of its mPR regulators. In total, glioblastoma expression profiles of 292 genes, including many known drivers of tumorigenesis and tumor subtype such as *RUNX1*, *PTPRN*, *FGFR3*, *TGFBR2*, and *DICER1*, were

significantly correlated ($p < 0.001$) with deletions at the loci of their mPR regulators. Strikingly, the expression profiles of more than half of these genes had stronger correlation with deletions at the loci of their mPR regulators than with deletions at their own loci.

To confirm functional mPR-based *PTEN* regulation by these 13 mPR regulators, we performed siRNA-mediated silencing of each gene in SNB19 cells, and measured the effect on *PTEN* using a *PTEN* 3' UTR luciferase reporter assays. Silencing of 11 of the 13 modulators in SNB19 lead to a significant ($p < 0.01$) reduction in *PTEN* luciferase activity, compared to negative controls (Figure 3.4(C)). To further validate that this regulatory mechanism is symmetric in nature, thus allowing *PTEN* expression to modulate the same 13 genes, we transfected SNB19 cells with *PTEN* 3' UTR and measured the effects on modulator expression (Figure 3.4(D)). This also addressed the potential issue that siRNA-mediated silencing may perturb endogenous miRs, possibly affecting the results displayed in Figure 3.4(C). Upregulation of 10 of the 13 modulators was significant ($p < 0.01$). Overall, 13/13 tested interactions were positive either in siRNA silencing or in 3' UTR expression assays. To ensure that the effects are not cell-line specific, we repeated the two experiments in the glioblastoma cell line SF188, using the subset of genes that were expressed in this cell line (9/13) (Figure 3.4(E),(F)). Results in SF188 confirmed the SNB19 results: indeed, silencing 9 of the 9 modulators lead to a significant reduction in *PTEN* luciferase activity, and transfection with *PTEN* 3' UTR upregulated 7 of the 9 modulators. Taken in aggregate, the cumulative Fisher's p-value across all the experiments is effectively below machine precision ($p < 1e\text{-}221$ based on an analytical estimate).

To show that these effects are specific to mPR regulators, six negative control genes were selected randomly among those that (a) are not *PTEN* neighbors in the mPR network, (b) have variable length UTRs, and (c) show a variety of correlation patterns with *PTEN*'s mRNA expression (positively correlated, negatively-correlated, and uncorrelated). Randomly selected genes include *TMEM149* (30-base 3' UTR), *POFUT1* (4003-base 3' UTR), *DDX24* (269-base 3' UTR), *SLC46A3* (1416-base 3' UTR), *EXTL3* (2819-base 3' UTR), *PIK3R2* (1239-base 3' UTR), and *EHMT2* (324-base 3' UTR). Of these, expression profiles of *DDX24* and *SLC46A3* are significantly positively correlated with that of *PTEN*, while *POFUT1* expression is significantly anti-correlated with *PTEN* expression. All of these genes, except for *POFUT1*, were highly expressed and could be silenced in SNB19 cells, while only *DDX24*, *EHMT2*, *EXTL3*, and *POFUT1* were expressed and could be silenced in SF188 cells. As predicted, *PTEN* 3' UTR luciferase activity was

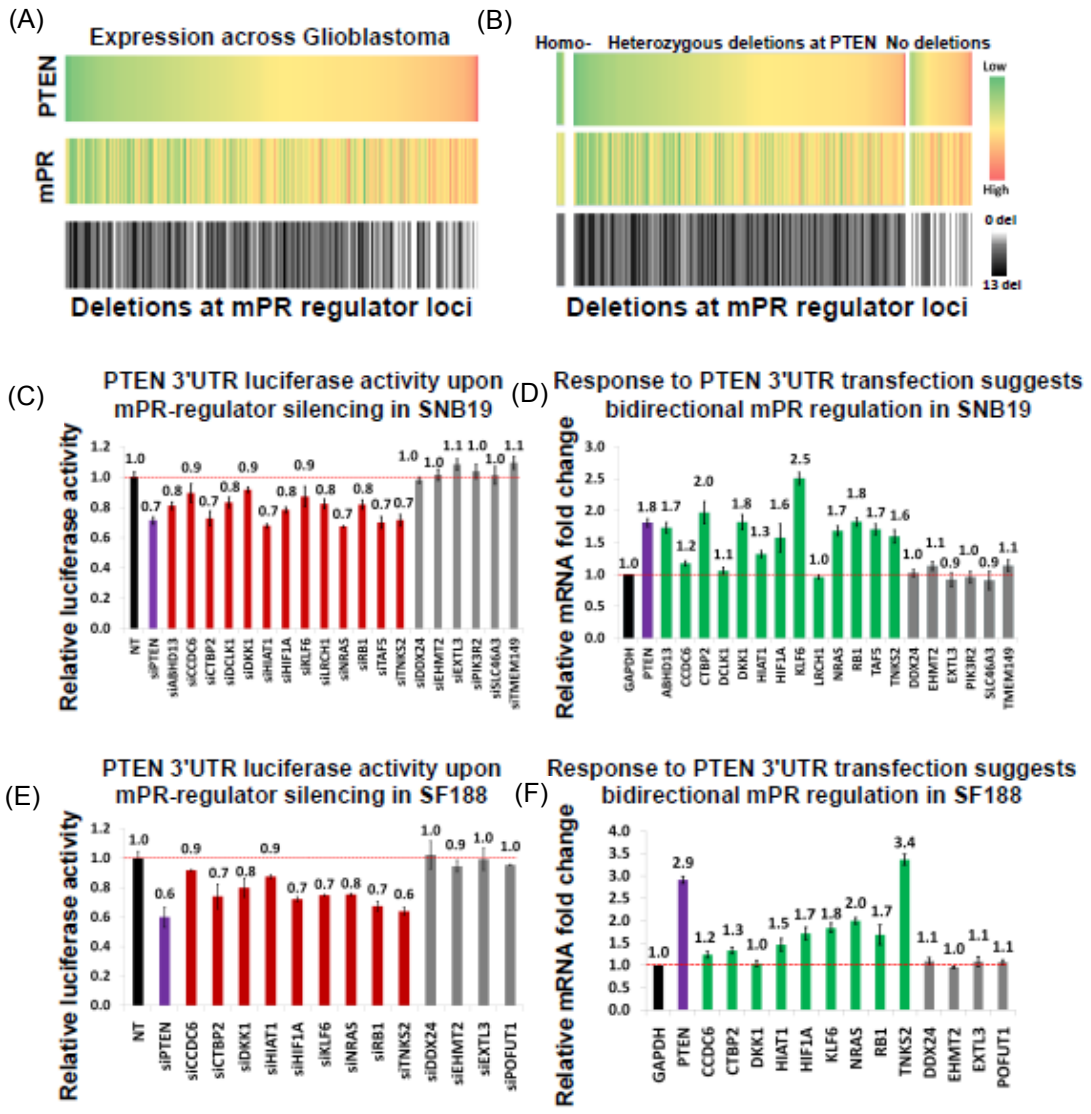


Figure 3.4. *PTEN* expression is correlated with the expression of its mPR regulators

(A) PTEN is targeted by >500 mPR regulators and its expression is correlated with both their total gene expression and with deletions at their loci; in aggregate, 97% of the TCGA glioma tumors have at least one deletion in a PTEN mPR-regulator locus. We selected 13 mPR regulators of PTEN with enriched locus deletions in *PTEN* intact tumors. As shown, their collective deletions and total expression are both significantly correlated with *PTEN* expression (p_D < 2e-10 and p_E < 5e-23, respectively). (B) Surprisingly, the correlation between *PTEN* and the aggregate expression across the 13 genes is significant in both samples with an intact *PTEN* locus and samples with heterozygous deletions (r_D = 0.40, p_D < 1e-09 and r_{WT} = 0.46, p_{WT} < 4e-04 by Pearson correlation, respectively). The range of *PTEN* expression in *PTEN* heterozygously deleted samples and in samples with an intact *PTEN* locus was virtually the same. (C) Individual siRNA mediated silencing of 13 PTEN mPR regulators reduced *PTEN* 3' UTR luciferase activity in SNB19 cells at 24h. Negative control targets (in grey) were unaffected. (D) Ectopic expression of *PTEN* 3' UTR increased expression of 13 PTEN mPR regulators in SNB19 cells at 24h, compared to empty vector. Negative control targets (in grey) were unaffected. (E),(F) Results in SNB19 were replicated in SNF188 cells for genes that are expressed in this cell line. Fold change was measured by qRT-PCR. Data are represented as mean ± SEM.

unchanged after silencing these genes in both cell lines, see Figure 3.4(C),(E). Furthermore, their mRNA expression was not significantly affected following transfection with *PTEN* 3' UTR, see Figure 3.4(D),(F). Note that the expression of these genes was affected by transfection with *PTEN* 3' UTR, see Figures 3.4(D),(F). Taken together, these results confirm cell-line independent, miR-mediated interactions between predicted mPR RNA pairs, including PTEN and its predicted mPR regulators, but not between these genes and other randomly selected genes (negative controls), regardless of their correlation with *PTEN* expression or the lengths of their UTRs.

3.7 Tumor growth is regulated by PTEN mPR interactions

To test whether PTEN mPR regulators may affect tumor cell growth, as previously shown for PTEN's post-transcriptional regulator *PTENP1* [7], we measured SNB19 and SF188 cell growth rates in response to transfection of PTEN cDNA (missing the 3' UTR) and 3' UTR, as well as to siRNA-mediated silencing of PTEN and of its Hermes-inferred mPR regulators (Figure 3.5). Transfection of *PTEN* 3' UTR upregulated the expression of its mPR neighbors, increased PTEN (protein) concentration, and reduced tumor cell growth rates. Conversely, siRNA-mediated silencing of 10/13 and 9/9 mPR-regulators reduced *PTEN* 3' UTR-luciferase expression and significantly accelerated SNB19 and SF188 cell growth, respectively. The effect of silencing these regulators was comparable to that of siRNA-mediated PTEN silencing, and the aggregate p-value for the significance of the increase in tumor cell growth computed by Fisher's method is below machine precision (i.e., $p \approx 0$).

3.8 Glioma regulators form dense sub-graph in the mPR network

The mPR network may explain significant crosstalk among different regulatory compartments of the cell that observed in perturbation experiments [24]. Indeed, further investigation of the mPR network revealed that known drivers of glioma tumorigenesis and glioblastoma subtypes *RB1*, *PTEN*, *RUNX1*, *PDGFRA*, *STAT3* and *VEGFA* [19,23,25] are part of a dense sub-graph of mutually mPR-interacting genes, see Figure 3.6(A). Ectopic expression of PTEN 3' UTR was effective in upregulating expression of the other genes in this sub-graph, while siRNA-mediated silencing of DICER and DROSHA (necessary for miR processing) was sufficient to abrogate the effect, suggesting that these interactions are miR-mediated, see Figure 3.6(B). To further confirm symmetric post- transcriptional regulation across all genes in the sub-graph, we

measured their response to transfection with the 3' UTRs of all other genes in the sub-network in SNB19 cells by qRT-PCR, except for VEGFA, whose 3' UTR cloning was not successful. Results confirmed that ectopic expression of the 3' UTRs of genes in this sub-network upregulated expression of the other genes (Figure 3.6(B),(C)). In particular, ectopic expression of *PTEN* and *RB1* 3' UTRs led to a >50% up-regulation of both genes, suggesting significant miR-mediated crosstalk between the *PTEN* and RB1 pathways, both implicated in gliomagenesis. Moreover, co-ectopic expression of 3' UTR pairs, at 50% concentration for each UTR, intensified the regulatory response (Figure 3.6(D)), suggesting that the effect of multiple mPR modulation is more than additive, as suggested by results shown in Figure 3.3(B).

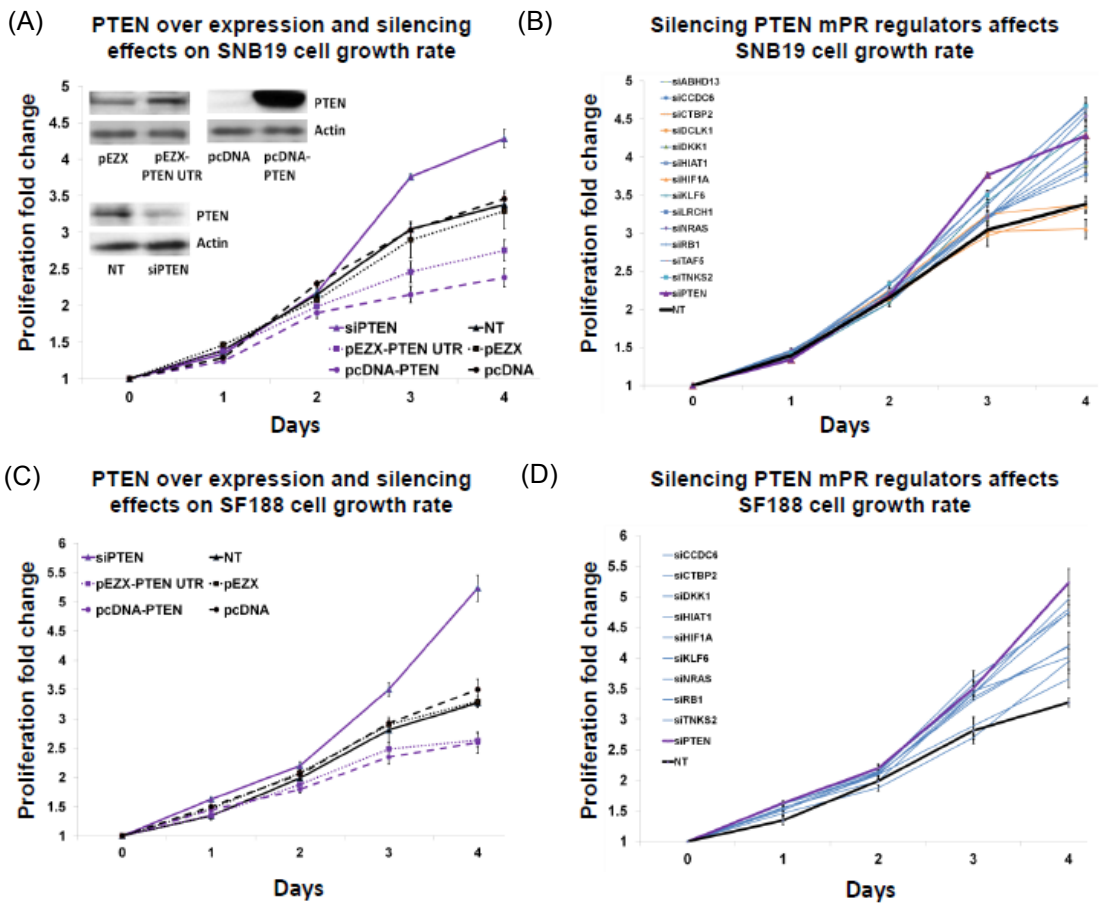


Figure 3.5. Silencing of PTEN mPR regulators accelerates tumor cell growth

(A) Cell proliferation assays were performed at 24h intervals, up to 4 days, following siRNA mediated PTEN silencing, *PTEN* cDNA ectopic expression, and *PTEN* 3' UTR ectopic expression. Protein levels of *PTEN* were assessed by Western blotting at day 1. (B) Cell proliferation assays were performed at 24h intervals, up to 4 days, following siRNA mediated silencing of 13 PTEN mPR regulators. Non-target (NT) siRNA was used as a control. (C),(D) Results in SNB19 were replicated in SNF188 cells for genes that are expressed in this cell line. Data are represented as mean ± SEM.

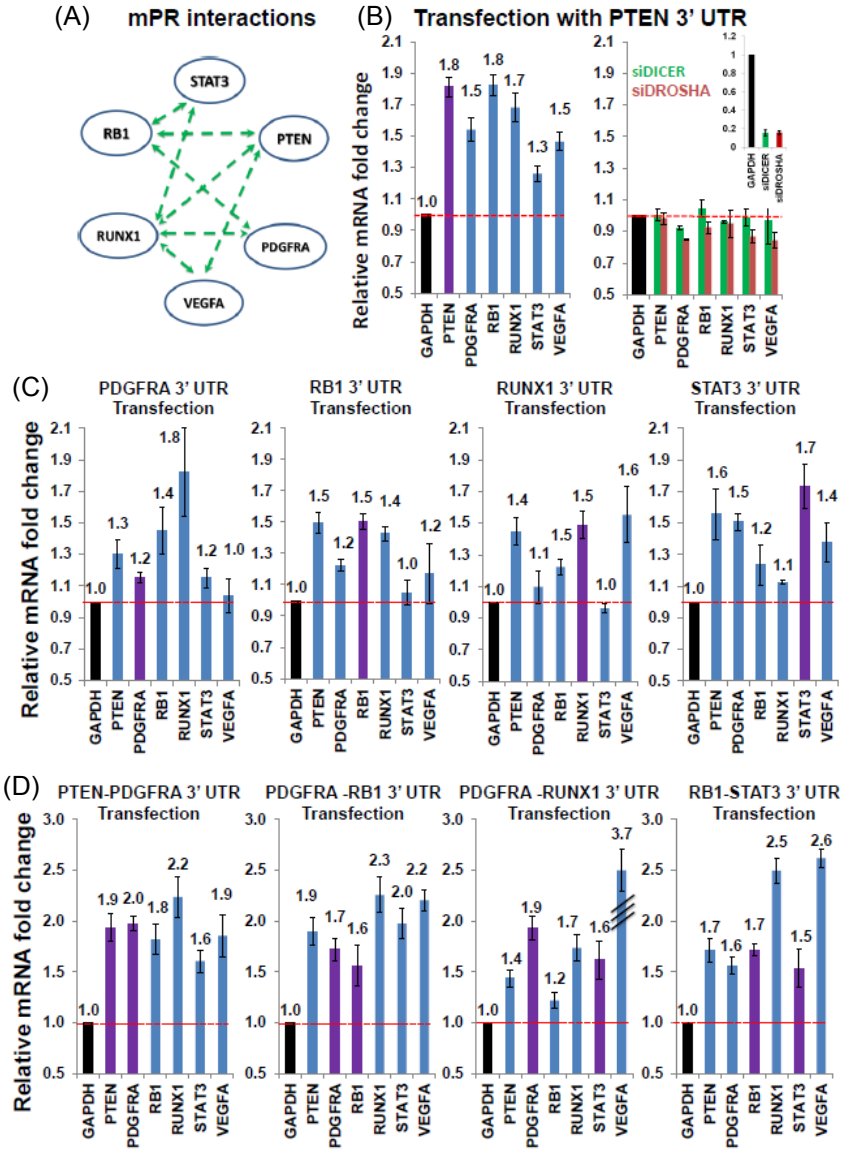


Figure 3.6. 3' UTR transfections confirm miR-mediated interactions between key drivers of glioma

(A) A tightly interconnected mPR network subgraph was identified, which includes established drivers of gliomagenesis. Sponge-mediated interactions inferred by Hermes are shown as dotted green lines. (B) Gene expression fold change of *PTEN*, *PDGFRA*, *RB1*, *RUNX1*, *STAT3*, and *VEGFA* at 24h following ectopic expression of *PTEN* 3' UTR, compared to an empty vector, with (right panel) and without (left panel) siRNA mediated silencing of DICER and DROSHA. (C) Gene expression fold change of *PTEN*, *PDGFRA*, *RB1*, *RUNX1*, *STAT3*, and *VEGFA* at 24h following ectopic expression of *PDGFRA*, *RB1*, *RUNX1* and *STAT3* 3' UTRs, compared to empty vector. (D) Gene expression fold change of *PTEN*, *PDGFRA*, *RB1*, *RUNX1*, *STAT3*, and *VEGFA* at 24h following ectopic expression of 3' UTR pairs, including double transfections of *PTEN* and *PDGFRA*, *PDGFRA* and *RB1*, *PDGFRA* and *STAT3*, and *RB1* and *STAT3* 3' UTRs. Gene expression was assessed by qRT-PCR. To highlight the significance of the change, Note that Y-axes start at 0.5 to better visualize the ratio between the experimental error and the change in expression. Data are represented as mean ± SEM.

Cooperativity between some of the regulators in the sub-network has been implicated in high-grade gliomagenesis [24], despite their distinct functions, lack of common transcriptional regulation, and large genomic distances. In particular, the loci of tumor suppressors PTEN and RB1 are frequently deleted in high-grade glioma (*PTEN*: 80%; RB1: 33%; *PTEN*+RB1: 85%). Analysis of 3' UTR luciferase activity of five of the six genes (Figure 3.6(A)), following siRNA-mediated silencing of PTEN and RB1, confirmed the presence of the predicted regulatory interactions (Figure 3.7). Thus, our results suggest that PTEN and RB1 regulate one another post-transcriptionally through 32 miRs in a common program, and that their availability significantly affects expression of other genes in the same sub-graph with an established role in gliomagenesis. Overall, 6 of 8 predicted interactions were confirmed by these assays. Additionally, of 11 experimentally validated interactions between these genes, 6 were predicted, suggesting a false negative rate of ~45%.

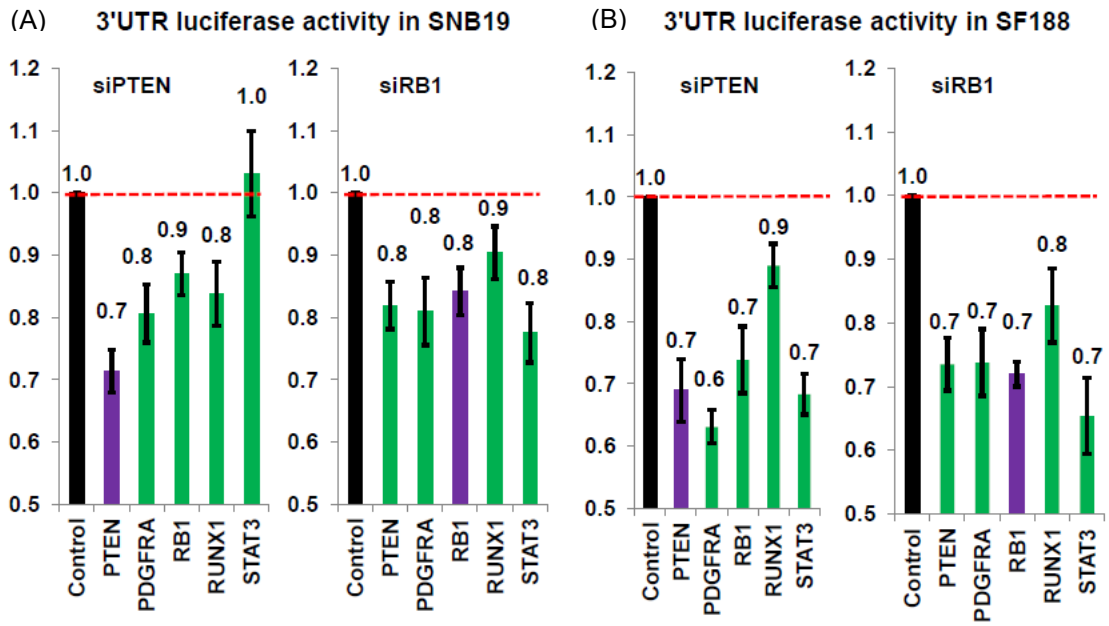


Figure 3.7. 3' UTR luciferase activity assays confirm miR-mediated interactions between key drivers of glioma

(A) 3' UTR luciferase activity of *PTEN*, *PDGFRA*, *RB1*, *RUNX1* and *STAT3* were measured in SNB19 cells at 24h following siRNA mediated silencing of PTEN and RB1 compared to non-targeting siRNAs (NT5) as control (in black). (B) Results in SNB19 were replicated in SNF188 cells. Similar to Figure 3.6, Y-axes start at 0.5. Data are represented as mean ± SEM.

3.9 Discussion

3.9.1 Hermes unveils an extensive layer of miR-mediated post-transcriptional regulation

Genome-wide Hermes analysis supports the existence of a miR-mediated, post-transcriptional regulation layer of unsuspected magnitude, the mPR network, effected by sponge interactions. While the specific mechanism of sponge modulation and the potential for miR-gene interactions were previously reported [6-8,19,26], both the extent and the functional relevance of this regulatory layer was unknown. In terms of size, the mPR layer rivals transcriptional regulation, supporting regulation of thousands of RNA species and modulating crosstalk between distinct regulatory pathways. Changes in two or more mPR regulators of a target gene may have effects comparable to transcriptional regulation (i.e. > 2-fold changes), as suggested by Figure 3.3(C) and shown in Figure 3.6(C) and Figure 3.6(D). The mPR network is implemented by sponge-mediated interactions that are generally symmetric in nature. A key and potentially confusing point is that our analysis suggests that mPR sponge interactions are mediated by relatively large miR programs, including on average 18 and up to 153 miRs. As a result, the effect of individual miRs is relatively negligible and mPR regulation is unlikely to be significantly affected by modulation of individual miRs or miR binding sites in isolation.

Importantly, while we have validated a substantial set of miR-mediated PTEN modulators in multiple cell lines, this by no means constitutes a thorough validation of the entire network. Yet, out of 28 experimentally validated interactions, all but 6 were confirmed (all but 3 if one considers both 3' UTR expression and siRNA mediated silencing assays in Figure 3.4(C) and 3.4(D). This suggests that false positive rates should be low (~10%-20%), comparing favorably with false positive rates in typical high-throughput experimental procedures. Thus, if globally validated using the experimental assays proposed in this manuscript, which is not currently feasible even using high-throughput approaches, a substantial number of the predicted interactions should be confirmed. Furthermore, we validated all of the pairwise interactions in the dense sub-graph that includes *PTEN*, *STAT3*, *VEGFA*, *PDGFRA*, *RUNX1*, and *RB1*. Of the 11 that were experimentally validated, 5 were not predicted, suggesting a false negative rate of ~45%, which is also competitive with experimental false negative rates.

3.9.2 MiR-activity modulators regulate pathogenesis of disease

It is important to note that while individual miR-mediated interactions may be weak, their regulatory effect in combination is substantial, see Figures 3.3(B) and 3.6(D). Furthermore, their ability to affect cellular phenotype is also significant and comparable to what was previously described for *PTENP1* [7], whose deletion was shown to be tumorigenic *in vivo*. This suggest that miR-mediated interactions between genes may play an important role in disease initiation and progression when dysregulated. Indeed, analysis of large glioblastoma datasets revealed that miR-mediated PTEN regulators are highly predictive of PTEN downregulation even when the *PTEN* locus is intact and may account for a significant proportion of the missing genetic variability of the *PTEN* locus.

In this study, we focused on *PTEN* as a key driver of gliomagenesis whose locus is often altered in glioblastoma samples [23]. However, regulation by miR-activity modulators is not limited to *PTEN* or to glioma. In addition, we showed that a variety of well-established drivers of tumorigenesis and tumor subtype in glioblastoma are regulated by miR-activity modulators, and our computational predictions suggest that other established oncogenes and tumor suppressors are similarly regulated. Since these effects are miR mediated and miR expression is strongly cell-context dependent, mPR networks are likely to be context-specific and their structure and contribution to disease initiation and progression will need to be studied independently in different contexts.

3.9.3 Direct screening methods are required for systematic prediction

Hermes, the algorithm used for the identification of miR-activity modulators, presents one key advantage. While it may be possible to infer sponge modulators by miR-target analysis alone, for instance by identifying genes whose transcripts share common miR binding sites, identification of functional miR targets is still largely inaccurate, with different methods predicting widely different interactions. Hermes circumvents this problem by first integrating evidence from multiple miRs in a common program and then by requiring direct, multivariate expression-based evidence for the predicted interaction, by conditional mutual information analysis. Thus, false negative predictions by miR-target prediction algorithms are much less critical than false positive predictions, as the latter dramatically reduce the statistical power of the method by increasing the number of hypotheses tested by the algorithm. On the other hand, even if miR program size is reduced

by false negatives, conditional mutual information analysis can still filter false positive interactions. As a result, rather than relying on existing algorithms for miR target prediction, which still have substantial false positive rates, we implemented Cupid specifically to reduce false positive predictions even if at the expense of some false negative predictions, see Experimental Procedures. Indeed, Cupid predicts fewer miR-target interactions than the intersection of three established algorithm TargetScan [27], PITA [28], and miRanda [29]. However, when we replaced Cupid predictions by the intersection of the three algorithms, 25 out of 26 experimentally validated mPR interactions in this study were missed. As a result, while our analysis does not suggest that Cupid may outperform other algorithms in terms of miR target identification, its specific design, aimed at minimizing false positives at the expense of false negatives, is uniquely tailored to inferring miR programs for further Hermes analysis.

3.10 Conclusion

Periodically, we are faced with the emergence of new regulatory layers, the post-transcriptional and histone modification ones being the latest additions. Every time this happens, we discover that these layers account for a significant amount of missing genetic and epigenetic variability in the etiology of disease. As a result, as suggested by our data, it is reasonable to expect that this novel and extensive miR-mediated interaction layer, which allows gene regulation without direct transcriptional or even post-transcriptional interactions, will also provide a number of clues to the dysregulation of key mechanisms of pathogenesis as well as to the regulation of normal cell physiology.

3.11 Experimental Procedures

We used a miR-activity modulator screening algorithm, *Hermes,* to identify candidate miR-activity modulators by finding genes whose expression is correlated with deviations in co-expression between miR programs and their targets using conditional mutual information. We used an integrative miR-target prediction algorithm, *Cupid*, to predict miR-target interactions and to assemble miR-regulatory programs for 3' UTRs. We identified genomic alterations using snapCGH [30]. Level 3 Agilent gene and miR expression data for glioma tumors were obtained from TCGA [17]. The glioblastoma cell lines SNB19 and SF188 were cultured under standard conditions. Transient transfections of expression vectors were used to over-express genes and 3' UTRs; siRNAs were used for mRNA silencing; real-time PCR, luciferase activity,

western blots, and proliferation assays were performed according to standard protocols. Our methods and experimental procedures are described in detail below.

3.11.1 Screening for miR-activity modulators

Similar to MINDY, which uses a CMI estimator (ΔI) to identify modulators of transcription factor activity, Hermes identifies candidate miR-activity modulators by finding genes whose expression is associated with changes in mutual information between regulators (miRs) and their targets. Both MINDY and Hermes rely on the idea that high CMI implies that modulator expression M is predictive of changes in regulatory activity of a regulator R on its target(s) T. However, while MINDY evaluates (M, R, T) triplets individually, Hermes evaluates them in sets $(M, \Pi_{miR}(T), T)$, based on a potentially large Cupid-inferred miR program $\Pi_{miR}(T)$ that targets T. Specifically, for each $miR_k \in \Pi_{miR}(T)$ in the program, Hermes evaluates the statistical significance (*p*-value) of the test $I[miR_k; T|M] > I[miR_k; T]$, where the variables indicate the expression of the corresponding RNA species. The CMI is estimated using an adaptive partitioning algorithm [31] by first iteratively partitioning the 3-dimentional expression space evenly into 8 partitions per iteration until partitions are balanced (p>0.05 by chi squared test), and then summing up MI across partitions. *P*-values for each triplet are computed using a null-hypothesis where the candidate modulator's expression is shuffled 1,000 times, thus preserving the pairwise mutual information between miR and target. Final significance across the entire program is then computed by converting the individual *p*-values, p_k, for each miR_k, to a X^2 test statistic using Fisher's method, where $X^2 = -2\sum_{k=1}^{N} \ln(p_k)$ and N is the total number of miRs in the program.

To avoid confusion, since modulators and targets are interchangeable when considering sponge-mediated interactions, we will replace M and T in the previous formalism with T_1 and T_2. When determining the existence of a $T_1 \leftrightarrow T_2$ sponge-mediated interaction between two genes, Hermes uses two distinct tests. First, the size of their common miR program, $\Pi_{miR}(T_1; T_2) = \Pi_{miR}(T_1) \cap \Pi_{miR}(T_2)$, is required to be statistically significant relative to the two individual miR programs size (FDR<*1e-02* by Fisher's exact test). Second, *p*-values for $I[\Pi_{mirs}(T_1; T_2); T_2|T_1] > I[\Pi_{mirs}(T_1; T_2); T_2]$ and $I[\Pi_{mirs}(T_1; T_2); T_1|T_2] > I[\Pi_{mirs}(T_1; T_2); T_1]$ are combined using Fisher's method to evaluate the global statistical significance of the interaction. Overall, fewer than 20% of the candidate interactions with significant common miR programs

passed the second test (at FDR<*1e-04*) and were included in the final mPR network. Moreover, only 0.02% of the candidate interactions with a significant common miR program, was statistically significant only in one direction.

We note that node connectivity in the mPR network is scale free and approximates $0.44x^{-1.2}$ with $R^2 = 0.95$. Node connectivity is strongly correlated ($r = 0.65$, by Pearson correlation) with the number of miRs that are predicted to target its corresponding gene. To evaluate the significance of the size of the mPR network, we generated random networks from permuted Cupid associations, while maintaining the distribution of regulator-target counts per gene and miR. Here, after permuting Cupid associations, 10 randomized mPR networks were generated. The sizes of these networks were normally distributed with a mean of $21{,}000 \pm 740$ interactions. Our results suggest that even when maintaining the node-connectivity distribution of the Cupid miR-target network, which is highly correlated with node connectivity in the mPR network, the size of the actual mPR network is significantly (>300 standard deviations away) larger than expected by chance. In addition, we note that mPR networks obtained from randomizing Cupid interactions were not scale free.

3.11.2 MiR-target interaction prediction

We used an integrative miR-target prediction algorithm, *Cupid*, to predict miR targets and to assemble miR-regulatory programs for 3' UTRs of interest. Cupid scores miR-binding sites by integrating (a) predicted site scores in RefSeq-annotated [32] 3' UTRs (December, 2010) from TargetScan, PITA and miRanda, (b) 46-vertebrate genome cross-species conservation scores by PhasCons [33], and (c) positional information relative to the 3' UTR start site. Cupid trains an SVM classifier using LIBSVM [34] to produce scores from 0 to 1 for each site that is predicted by at least one of the three algorithms by training against 684 validated miR targets obtained from miRecords [35] as of June, 2010. Site scores are then summarized using an array of summary functions that model miR binding-site interactions both linearly and non-linearly, and an SVM is then used to generate miR-target interaction scores using site scores and their summaries. Comparisons between We note that the majority of (70%) Cupid-predicted interactions are supported by TargetScan-predicted miR binding sites, but only 15% of candidate interactions supported by TargetScan sites are predicted by Cupid.

Cupid is specifically designed to produce very low false-positive rates, possibly at the expense of false-negative rates, by using stringent, integrative selection criteria. Interestingly, the number of Cupid-predicted interactions (486,100) is smaller than the total number of predictions that are common to all three algorithms (734,594). A 50% increase in the number of predicted miR-target interactions, when using the intersection of all three algorithms, would be expected to increase the size of the predicted mPR network. Instead, the opposite occurs and 25 out of the 26 miR-mediated interactions validated in this study are missed. Since Cupid is not used to predict individual miR interactions but rather to predict the presence of significant common miR regulation programs between two RNAs, characterization of its performance with respect to individual miR-target predictions is beyond the scope of this study.

3.11.3 Genomic alteration prediction

Genomic alterations were identified using snapCGH with 0.7 and 1.7 copy number cutoffs for identifying potential homozygous and heterozygous genomic deletions respectively. Normalized copy number was estimated as two times the snapCGH ratio [23].

3.11.4 Cell and culture condition

The glioma-derived cell lines SNB19 and SF188 were grown in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% FBS (Gibco/BRL). Freshly trypsinized cells were suspended at 3×10^5 cells/ml in standard culture medium and seeded at a density of 3×10^5 cells per well in standard six-well tissue culture plates. After seeding, the cells were incubated at 37°C in a 95% air/5% CO_2 humidified atmosphere, and one milliliter of fresh medium was supplied every other day to the cultures after removal of the supernatant.

3.11.5 RNA interference and reverse transfection

Silencer® select non-targeting siRNA (NT5), validated and pre-designed siRNAs targeting human *B2M, DICER1, DROSHA, PALB2*, and *WIPF2* were purchased from Ambion. siGONOME non-targeting siRNA pools and SMARTpool siRNAs targeting human *ABHD13, BMI1, CCDC6, CTBP2, DCLK1, DDX24, DKK1, EHMT2, EXTL3, HIAT1, HIF1A, KLF6, LRCH1, NRAS, PIK3R2, PTEN, RB1, RUNX1, SLC46A3, TAF5, TMEM149*, and *TNKS2* were purchased from Dharmacon. Reverse transfection of siRNA was performed with the transfection reagent, Lipofectamine RNAiMAX (Invitrogen), following the manufacturer’s protocol.

In general, the siRNA was diluted in serum and antibiotic-free Opti-MEM (Invitrogen) and then mixed with the transfection reagent, Lipofectamine RNAiMAX. The mixture of siRNA andLipofectamine RNAiMAX in Opti-MEM was then added into the plates and incubated at room temperature for 20 min. Cells suspended in antibiotic-free medium were counted and plated into the plates at the same cell number per well. The cells were then incubated at 37 °C for 24 h. Titration of the siRNA and the transfection reagent was performed (not shown), and the lowest working amounts of the siRNA and the transfection reagent were applied in the present study.

3.11.6 Over-expression and forward transfection

The *WNT7A* plasmid, pCMV6-XL4-WNT7A, the empty vector, pCMV6-XL4, and the firefly luciferase reporter gene followed by *RUNX1* 3' UTR in pMirTarget vector were purchased from Origene. The firefly luciferase reporter gene followed by *PTEN* 3' UTR in pEZX vector was purchased from GeneCopoeia. 3' UTRs of *PDGFRA* (NM_006206.4), *RB1* (NM_000321.2), and *STAT3* (NM_139276.2) were amplified from normal human genomic DNA by PCR using specific primers bearing MluI/PmeI restriction sites. Amplified products (RB1 1950bp) (STAT3 2345bp) (PDGFRA 2975 bp) were cloned into the MluI/PmeI sites of pMIR-REPORTTM vector (AMBION #AM5795) and sequence verified. *RB1* primers: forward 5'-aaacgcgtTCAGAGCGGAGAAAGCAT-3', reverse 5'-aagtttaaacACTGCACTAGAGACAAAGACG-3'. *STAT3* primers: forward 5'-aaacgcgtACCTTTGACATGGAGTTGAC-3', reverse 5'-aagtttaaacCATTGGAATTTGAATGCAG-3'. *PDGFRA* primers: forward 5'-aaacgcgtAGACCATTGAAGACATCGAC-3', reverse 5'-aagtttaaacGGGCATTCGTAATACATTTT-3'. Forward transfection of the plasmid was performed with the transfection reagent, Lipofectamine 2000 (Invitrogen), following manufacturer protocol. In general, cells attached to the culturing surface were washed with phosphate-buffered saline, and the medium was replaced with 1 ml of Opti-MEM with 2% fetal bovine serum. Two micrograms per well in a 6-well plate of the plasmid was then mixed with 5 µl/well of Lipofectamine 2000 in Opti-MEM and 20 min later the mixture was added to the wells. Double transfections were achieved using half the plasmid quantity. After 6 hours of transfection, the cells were then cultured in regular medium for 24 h and subsequently harvested.

3.11.7 Real-time quantitative RT-PCR analysis

Total RNA was extracted from cells with the RNeasy mini kit (Qiagen, Valencia, CA) and depleted of contaminating DNA with RNase-free DNase (Qiagen). Equal amounts of total RNA (1 µg) were reverse-transcribed using qScript™ cDNA Synthesis kit (Quanta Biosciences). The first-strand cDNA was used as a template. Real-time PCR was carried out using SYBR green fluorescence. Two microlitres of RT were used in a 25-ul reaction. Each sample was assayed in three independent RT reactions and triplicate reactions were performed and normalized to the *GAPDH* expression levels. Negative controls included the absence of enzyme in the RT reaction and the absence of template during PCR. Relative quantification of gene expression was performed with the comparative CT method. Primers used for quantitative RT-PCR analyses were synthesized by Sigma-Aldrich. *PTEN* primers: forward 5'-TCCCTCAGCCGTTACCTGTGTGT-3', reverse 5'-TCTGAGGTTTCCTCTGGTCCTGGT-3'. *STAT3* primers: forward 5'-CGGCCTCTGCCGGAGAAACA-3', reverse 5'-TCCAAGGGGCCAGAAACTGCC-3'. *VEGFA* primers: forward 5'-GAGGGCCTGGAGTGTGTGCC-3', reverse 5'-GCTCACCGCCTCGGCTTGTC-3'. *RUNX1* primers: forward 5'-ACCACAGGGTTTCGCAGCGT-3', reverse 5'-CGGTGGAAGGCGGCGTGAAG-3'. *PDGFRA* primers: forward 5'-GAAGGCACGCCGCTTCCTGAT-3', reverse 5'-ACACGGCCCTCCACGGTACT-3'. *RB1* primers: forward 5'-TGGCGTGCGCTCTTGAGGTT-3', reverse 5'-AGAGCCATGCAAGGGATTCCATGA-3'. *TNKS2* primers: forward 5'-ACGGCGGGCAGGAAATCCAC-3', reverse 5'-TCGGATGGTTGGCTCAGCTCCA-3'. *CTBP2* primers: forward 5'-GGACCGAACCGGGAGCCATG-3', reverse 5'-TGCGTGCATGACGCCACTATGA-3'. *NRAS* primers: forward 5'-ACATGAGGACAGGCGAAGGCTT-3', reverse 5'-TGGCCAGTTCGTGGGCTTGTTT-3'. *TAF5* primers: forward 5'-TTGGGCCGGACTGCTTACCCT-3', reverse 5'-TCCGTAGACAGGCCCACTGTGA-3'. *HIAT1* primers: forward 5'-GGGACCGGCCCTCTATGGATTCA-3', reverse 5'-AAGGGAGGGCCAGGGATGATGG-3'. *WNT7A* primers: forward 5'-CCCGGGCGGGCTATGTTGATT-3', reverse 5'-GCTTGCGCCCAGAGCTACCA-3'. WIPF2 primers: forward 5'-CAGCCCGAGACCCTCCCAGT-3', reverse 5'-GCCCAGCTGGCGTCCTTGA-3'. *PALB2* primers: forward 5'-TCTGTCGCCTGCCCGATGGA-3', reverse 5'-CGCTGAAGGCGGGCTAGTGT-3'. LMO3 primers: forward 5'-TGTGGCTCAGATGCGGTCAACAC-3', reverse 5'-TTTCGGTTGCAGCCAGCACAA-3'. *PRKACB* primers: forward 5'-TGCACGGTTCTATGCAGCTCAGA-3',

reverse 5’-ATGCCCACCAATCCACTGCCTT-3'. *ZNF236* primers: forward 5’-GAGCAGAGCCCTGCGCAACA-3’, reverse 5’-GAGCTGGGAGCCTGCAGCAA-3'. *ZNF238* primers: forward 5’-GTAACAGACCTGGAGCCAGCAGGAC-3’, reverse 5’-GAGCGAAAGCGGGGGCTGTAA-3'. *PAK7* primers: forward 5’-ACTGTCTTCTGGACCTCTGAGACCA-3’, reverse 5’-TGTGTTCAAAGTTGGACGGGCCA-3'. *ABHD13* primers: forward 5’-AGCACTTGCCATTGCAACAACCC-3’, reverse 5’-GCAGCCCTGTAGCAAAATTCAGAAC-3'. *DCLK1* primers: forward 5’-CCGAGGCACATCCCTGCACTAGT-3’, reverse 5’-CGCGACCCTCGGCTGTATCT-3'. *HIF1A* primers: forward 5’-AGCCCTAACGTGTTATCTGTCGCT-3’, reverse 5’-GCTGCATGATCGTCTGGCTGCT-3'. *LRCH1* primers: forward 5’-CCCACGGTCGGTTGCAAGCA-3’, reverse 5’-CCCTGGAGCGGAGGGCAGAA-3'. *GAPDH* primers: forward 5’-AACTTTGGTATCGTGGAAGGA-3’, reverse 5’-CAGTAGAGGCAGGGATGATGT-3'. The real-time PCR of mature miRNAs was performed using the TaqMan mature miRNA PCR assay kit from Invitrogen, following the manufacturer’s protocol.

3.11.8 Cell proliferation assay

Starting from Day 1 after gene silencing or over-expression, cell proliferation was measured each day for 4 constitutive days using the PrestoBlue™ Reagent (Invitrogen) according to the manufacturer’s protocol. In general, the PrestoBlue™ reagent was added directly to cells in a 96-well plate and incubated at 37°C for 20-60 minutes. The plate was then transferred to a fluorescence reader to measure signal.

3.11.9 Dual luciferase reporter assay

Twenty-four hours after gene silencing, firefly luciferase activity was measured and normalized by renilla luciferase activity. The cells were cultured in a 96-well plate. 24 hours after transfection, both firefly luciferase and Renilla luciferase activities were measured using the Luc-Pair™ miR Luciferase Assay Kit (GeneCopoeia). Firefly luciferase activity was then normalized with Renilla luciferase activities in the same well.

3.11.10 Statistical analysis

All experiments were performed at least in triplicate and representative results are shown. All data are shown as the mean ± standard error. Student's t-tests were used to evaluate statistical significances between different treatment groups.

References

1. Gabriely, G., Yi, M., Narayan, R.S., Niers, J.M., Wurdinger, T., Imitola, J., Ligon, K.L., Kesari, S., Esau, C., Stephens, R.M.*, et al.* (2011). Human Glioma Growth Is Controlled by MicroRNA-10b. Cancer Res.

2. Godlewski, J., Nowicki, M.O., Bronisz, A., Williams, S., Otsuki, A., Nuovo, G., Raychaudhury, A., Newton, H.B., Chiocca, E.A., and Lawler, S. (2008). Targeting of the Bmi-1 oncogene/stem cell renewal factor by microRNA-128 inhibits glioma proliferation and self-renewal. Cancer Res *68*, 9125-9130.

3. Kim, H., Huang, W., Jiang, X., Pennicooke, B., Park, P.J., and Johnson, M.D. (2010a). Integrative genome analysis reveals an oncomir/oncogene cluster regulating glioblastoma survivorship. Proc Natl Acad Sci U S A *107*, 2183-2188.

4. Kim, T.M., Huang, W., Park, R., Park, P.J., and Johnson, M.D. (2011). A Developmental Taxonomy of Glioblastoma Defined and Maintained by MicroRNAs. Cancer Res *71*, 3387-3399.

5. Kwak, H.J., Kim, Y.J., Chun, K.R., Woo, Y.M., Park, S.J., Jeong, J.A., Jo, S.H., Kim, T.H., Min, H.S., Chae, J.S.*, et al.* (2011). Downregulation of Spry2 by miR-21 triggers malignancy in human gliomas. Oncogene.

6. Krol, J., Loedige, I., and Filipowicz, W. (2010). The widespread regulation of microRNA biogenesis, function and decay. Nat Rev Genet *11*, 597-610.

7. Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W.J., and Pandolfi, P.P. (2010). A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature *465*, 1033-1038.

8. Arvey, A., Larsson, E., Sander, C., Leslie, C.S., and Marks, D.S. (2010). Target mRNA abundance dilutes microRNA and siRNA activity. Mol Syst Biol *6*, 363.

9. Ebert, M.S., Neilson, J.R., and Sharp, P.A. (2007). MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. Nat Methods *4*, 721-726.

10. Lee, D.Y., Jeyapalan, Z., Fang, L., Yang, J., Zhang, Y., Yee, A.Y., Li, M., Du, W.W., Shatseva, T., and Yang, B.B. (2010). Expression of versican 3'-untranslated region modulates endogenous microRNA functions. PLoS ONE *5*, e13599.

11. Jeyapalan, Z., Deng, Z., Shatseva, T., Fang, L., He, C., and Yang, B.B. (2011). Expression of CD44 3'-untranslated region regulates endogenous microRNA functions in tumorigenesis and angiogenesis. Nucleic Acids Res *39*, 3026-3041.

12. Wang, K., Saito, M., Bisikirska, B.C., Alvarez, M.J., Lim, W.K., Rajbhandari, P., Shen, Q., Nemenman, I., Basso, K., Margolin, A.A.*, et al.* (2009b). Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. Nat Biotechnol *27*, 829-839.

13. Wang, K., Alvarez, M.J., Bisikirska, B.C., Linding, R., Basso, K., Dalla Favera, R., and Califano, A. (2009a). Dissecting the interface between signaling and transcriptional regulation in human B cells. Pac Symp Biocomput, 264-275.

14. Zhao, X., D, D.A., Lim, W.K., Brahmachary, M., Carro, M.S., Ludwig, T., Cardo, C.C., Guillemot, F., Aldape, K., Califano, A.*, et al.* (2009). The N-Myc-DLL3 cascade is suppressed by the ubiquitin ligase Huwe1 to inhibit proliferation and promote neurogenesis in the developing brain. Dev Cell *17*, 210-221.

15. Wang, K., Banerjee, N., Margolin, A.A., Nemenman, I., and Califano, A. (2006). Genome-wide discovery of modulators of transcriptional interactions in human B lymphocytes. RECOMB 2006 also Lecture Notes in Computer Science *3909*, 348-362.

16. Fisher, R.A. (1925). Statistical methods for research workers (Edinburgh, London,, Oliver and Boyd).

17. TCGA-Consortium (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature *455*, 1061-1068.

18. Maslov, S., and Sneppen, K. (2002). Specificity and stability in topology of protein networks. Science *296*, 910-913.

19. Carro, M.S., Lim, W.K., Alvarez, M.J., Bollo, R.J., Zhao, X., Snyder, E.Y., Sulman, E.P., Anne, S.L., Doetsch, F., Colman, H.*, et al.* (2010). The transcriptional network for mesenchymal transformation of brain tumours. Nature *463*, 318-325.

20. Barrat, A., Barthelemy, M., and Vespignani, A. (2008). Dynamical processes on complex networks (Cambridge, UK ; New York, Cambridge University Press).

21. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res *19*, 1639-1645.

22. Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res *28*, 27-30.

23. Verhaak, R.G., Hoadley, K.A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M.D., Miller, C.R., Ding, L., Golub, T., Mesirov, J.P.*, et al.* (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell *17*, 98-110.

24. Chow, L.M., Endersby, R., Zhu, X., Rankin, S., Qu, C., Zhang, J., Broniscer, A., Ellison, D.W., and Baker, S.J. (2011). Cooperativity within and among Pten, p53, and Rb pathways induces high-grade astrocytoma in adult brain. Cancer Cell *19*, 305-316.

25. Parsons, D.W., Jones, S., Zhang, X., Lin, J.C., Leary, R.J., Angenendt, P., Mankoo, P., Carter, H., Siu, I.M., Gallia, G.L.*, et al.* (2008). An integrated genomic analysis of human glioblastoma multiforme. Science *321*, 1807-1812.

26. Su, W.L., Kleinhanz, R.R., and Schadt, E.E. (2011). Characterizing the role of miRNAs within gene regulatory networks using integrative genomics techniques. Mol Syst Biol *7*, 490.

27. Lewis, B.P., Burge, C.B., and Bartel, D.P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell *120*, 15-20.

28. Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007). The role of site accessibility in microRNA target recognition. Nat Genet *39*, 1278-1284.

29. Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D.S. (2003). MicroRNA targets in Drosophila. Genome Biol *5*, R1.

30. Marioni, J.C., Thorne, N.P., and Tavare, S. (2006). BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. Bioinformatics *22*, 1144-1146.

31. Darbellay , G., and Vajda, I. (1999). Estimation of the Information by an Adaptive Partitioning of the Observation Space. IEEE Trans on Information Theory *45*, 1315--1321.

32. Pruitt, K., Tatusova, T., and Maglott, D. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res *33*, D501 - 504.

33. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S.*, et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res *15*, 1034-1050.

34. Fan, R., Chen, P., and Lin, C. (2005). Working set selection using the second order information for training SVM. Journal of Machine Learning Research *6*, 1889-1918.

35. Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., and Li, T. (2009). miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res *37*, D105-110.

Chapter 4: Constructing the pan-cancer ceRNA network across multiple tumors

4.1 Introduction

Through my analysis of four distinct tumor-specific ceRNA networks, I identified more than 160,000 interactions that are common to all four tumor contexts, including of glioblastoma, the carcinoma of breast, prostate, and ovary. Genome-wide statistical analysis and targeted experimental assays confirmed these interactions in a dozen additional cellular contexts, both tumor and non-tumor related, thus further supporting their predicted *pan-cancer* nature.

4.2 Assembly of ceRNETs

To dissect the contribution of individual miRNA mediators, I extended the previously published Hermes algorithm [1] for the reverse engineering of ceRNA networks (ceRNETs) to also identify the miRNA participating in each predicted interaction. Hermes predicts ceRNA interactions by identifying RNA-pairs with (a) a sufficiently large set of shared miRNA species (miRNA program), and (b) a statistically significant conditional mutual information between the expression profiles of the miRNA program and each candidate ceRNA, given the expression of the other candidate ceRNA. The latter can be effectively computed from a large set of miRNA and mRNA expression profiles from the same samples. The method is detailed below. Hermes predicts ceRNA interactions based on the relative size of shared miRNA regulatory programs between two genes based on predictions by Cupid [1], and the conditional mutual information between these genes and their shared miRNA program. Namely, given genes T_i and T_j, and the set of miRNAs that regulate them $\Pi_{miR}(T_i)$ and $\Pi_{miR}(T_j)$, their shared program is identified by taking the intersection $\Pi_{miR}(T_i; T_j) = \Pi_{miR}(T_i) \cap \Pi_{miR}(T_j)$. First, Hermes tests that the size of $\Pi_{miR}(T_i, T_j)$ relative to the sizes of the individual programs is statistically significant at FDR<1E-02 by Fisher's exact test. Then, Hermes evaluates the statistical significance p_{kij} (*p*-value) of the test $I[miR_k; T_i | T_j] > I[miR_k; T_i]$, where the variables indicate the expression of the corresponding RNA species. The CMI is estimated using an adaptive partitioning algorithm [2] by first iteratively partitioning the 3-dimentional expression space evenly into 8 partitions per iteration until partitions are balanced (p>0.05 by chi squared test), and then summing up CMI across partitions. *P*-values for each triplet are computed using a null-hypothesis where the candidate modulator's

expression (T_j) is shuffled 1,000 times, thus preserving the pairwise mutual information between miRNA and target. Final significance across the entire program is using Fisher's method to integrate both regulatory directions, i.e. T_i affecting miR_k regulation of T_j as well as T_j affecting miR_k regulation of T_j, for all the miRNAs in the shared miRNA program $\Pi_{miR}(T_i; T_j)$. Specifically, the value $X^2 = -2\sum_{k=1}^{N} \ln(p_{kij}p_{kji})$ is distributed as a X^2 distribution, with $4N$ degrees of freedom, where N is the number of miRNAs in the shared program. Finally, only prediction passing significance of FDR<*1e-03* were selected. Note that selected predictions by Hermes have been validated in glioblastoma cell lines [1].

In order to identify miRNA mediators in addition to ceRNA interactions, I modified the Hermes to perform greedy addition of miRNA mediators and to optimize the combined *p*-value for each predicted interaction. Namely, for each candidate interaction, I search for the minimum combined *p*-value through greedy forward inclusion of individual miRNAs. Mediators are included only if they improve the combined *p*-value as estimated using Fisher's method. Those that fail to improve the combined *p*-value lack functional evidence for mediating ceRNA regulation.

I used Hermes to construct ceRNETs for glioblastoma using gene and miRNA expression (423 samples, 12,032 genes, 469 miRNAs profiled) [3], ovarian cancer (583 samples, 12,032 genes, 713 miRNAs profiled) [4], prostate cancer (140 samples, 23,614 genes, 367 miRNAs profiled) [5] and breast cancer (207 samples, 18,748 genes, 524 miRNAs profiled) [6]. The resulting predicted ceRNETs had 527,430 (glioblastoma), 532,869 (ovarian), 476,456 (prostate) and 447,011 (breast) predicted interactions. The four ceRNETs, including the interacting ceRNA pairs and the associated miRNAs mediators. On average, fewer than 50% of candidate miRNA species shared by an interacting ceRNA pair (based on miRNA-target analysis) were found to functionally contribute to the interaction; see Figure 4.1. This is expected because only some of the miRNAs will be expressed in a range that may induce significant ceRNA coupling. Since miRNA expression is highly context specific [7], it is likely that different subsets of the shared miRNAs are expressed in a kinetically relevant range in each cellular context. Across the four tumors, however, seventeen miRNAs were shared on average by an interacting ceRNA-pair. Of these, seven were predicted to functionally mediate the interaction, on average (Figure 4.1). I thus expect that, the majority of Hermes-inferred ceRNA interactions in these tumors should induce significant coupling across tumor contexts.

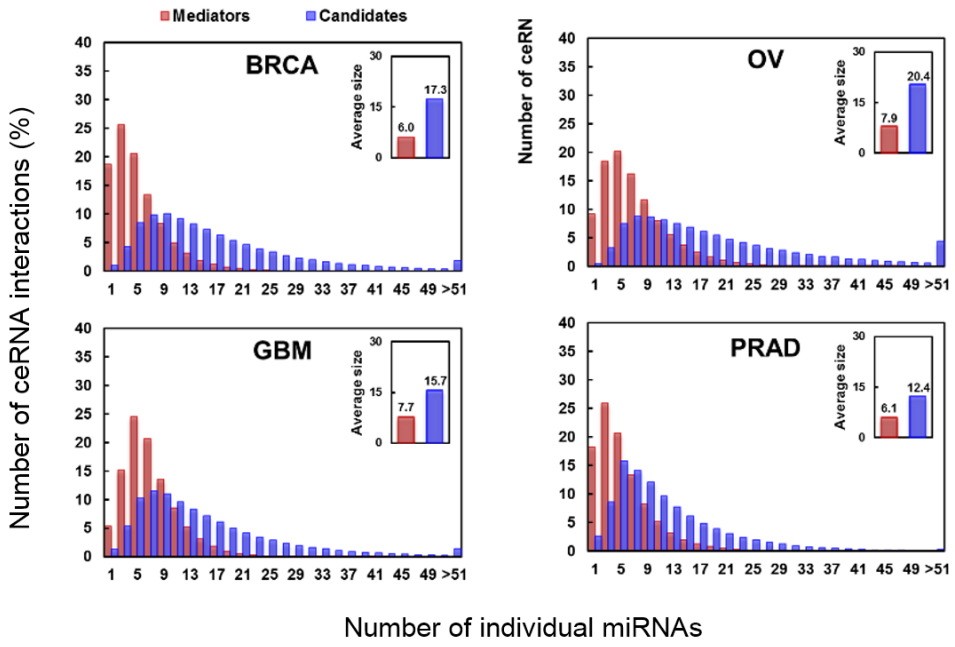


Figure 4.1. The distributions of candidate and inferred miRNA mediators in each of the four networks

Fewer than half of the candidates are predicted to consistently mediate ceRNA interactions in each context.

4.3 ceRNA interactions are ubiquitous across distinct tumor contexts

Considering the four ceRNETs for breast cancer, ovarian cancer, prostate cancer, and glioblastoma, an average of 298,570 Hermes-inferred interactions were overlapped (i.e., identically inferred) across any two tumor contexts. This represents approximately 43% of all Hermes-inferred ceRNA interactions in each context. This is a strikingly high fraction, especially considering that, given an estimated 20% false-negative rate [1], no more than 64% conserved interactions could be expected, even between two identical ceRNETs. Furthermore, most regulatory networks are highly context-specific. For instance, conservation between ARACNe-inferred transcriptional networks [8,9] in glioblastoma and breast cancer is just 1%, and conservation of protein-protein interactions by high-throughput experimental methods, such as yeast-2-hybrids, rarely exceeds >10% overlap [10]. When coupled with the fact that inferred ceRNA interactions had 17 mediating miRNAs on average, this result completely supports the kinetic model prediction that

ceRNA interactions mediated by many miRNA should be conserved and independent of cellular context. Even more striking, a total of 164,623 ceRNA interactions, about a third of those found in each individual tumor context at FDR<1E-3, were predicted to be frequently occurred across all four ceRNETs (Figure 4.2). These ceRNA interactions form an *pan-cancer ceRNA network* (i.e., the PC-ceRNET). To test the statistical significance of this finding, I performed permutation tests, where in each test I swapped 1,000,000 edges at random between ceRNA pairs in each ceRNET, thus preserving node connectivity and network topology and ensuring that each random interaction is supported by a realistic miRNA program [1]. Out of 10^{12} tests, I never observed a comparable number of conserved interactions, suggesting that the statistical significance is below 1E-12.

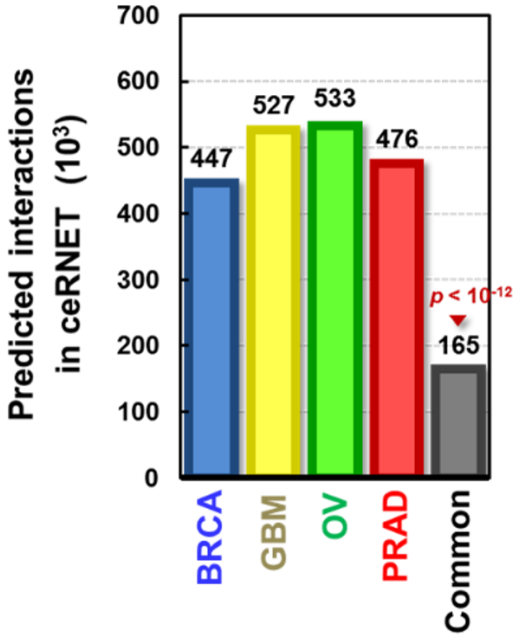


Figure 4.2. The overlap of ceRNETs in four tumor datasets identified pan-cancer ceRNA network (PC-ceRNET)

ceRNETs were inferred from RNA profile datasets from breast cancer, glioblastoma, ovarian cancer and prostate cancer tumors. (A) A significantly large set of nearly 165,000 interactions, which are common to all four networks (p < 1E-12).

To experimentally test the conservation of PC-ceRNET interactions I focused on two small subnetworks: one including the three oncogenes *CCND1*, *HIF1A* and *HMGA2* and the other the three tumor-suppressors *PTEN*, *RB1* and *TP53*. Predicted interactions between these ceRNAs were tested in A549 (lung cancer), C3A (liver cancer), HT-29 (colon cancer), SK-MEL-28 (melanoma), MCF7 (breast cancer), U2-OS

(osteosarcoma), PC3 (prostate cancer) and SK-OV-3 (ovarian carcinoma). These represent eight distinct tumor contexts, including five not originally used for the inference of these interactions. My colleagues measured the differential expression of each gene in the two sub-networks by qRT-PCR, following transfection of each 3' UTR, including their own 3'-UTR as a positive control. As negative controls, they used 3' UTRs of genes not predicted to target the six selected genes. Results are summarized in Figure 4.3. In total, in 80% of the experiments, predicted targets were significantly upregulated ($p \leq 0.05$, by T-test), compared to 4.5% of negative-control experiments. The difference is highly statistically significant ($p \leq$ 2E-40, by Fisher Exact Test). While significant and highly consistent, individual responses were relatively small, with an average 1.42-fold increase in gene expression, consistent with previously reported results [1].

Interestingly, while ceRNA interactions were conserved across tumor types, the miRNAs predicted to mediate them varied significantly between contexts. Indeed, systematic analysis suggested no statistically significant overlap in the miRNA repertoire that mediates same ceRNA interactions across contexts. For 66% of the PC-ceRNA interactions, no miRNA was found to functionally mediate a ceRNA interaction in all four ceRNETs; see the following section.

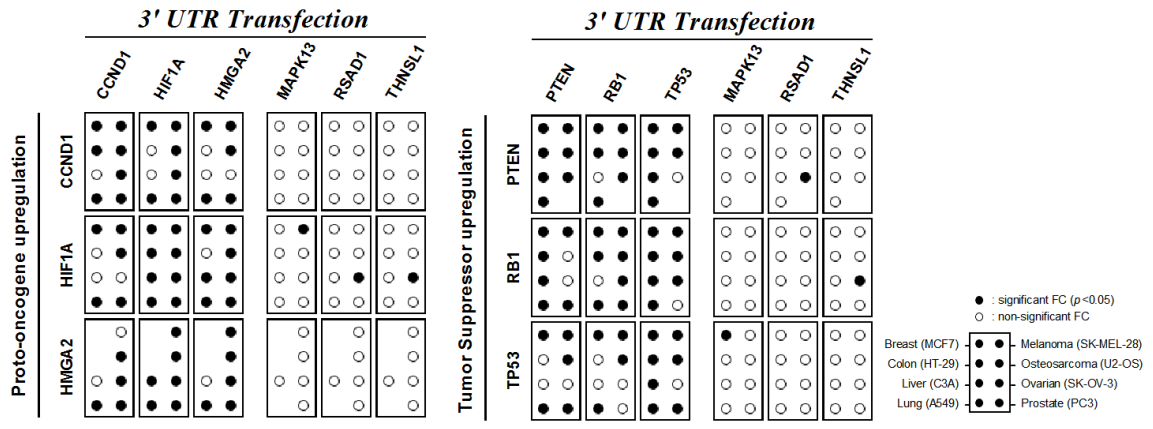


Figure 4.3. Validation of oncogene and tumor-suppressor subnetworks in the PC-ceRNET

Significant mRNA up regulation in response to 3' UTRs transfection of candidate ceRNA regulators in eight cell lines is shown for a three-oncogene subnetwork and for a three-tumor-suppressor subnetwork. Negative controls include *MAPK13*, *RSAD1* and *THNSL1*, which are not predicted to regulate genes in the two subnetworks. For each transfected 3' UTR and each target gene, a panel identifies the cell lines where targets were significantly ($p < 0.05$) upregulated in response to transfection as black dots in a matrix representing the eight tumor cell lines. Missing dots designate cell lines where target expression was too low to be measured.

4.4 Estimating the conservation of PC-ceRNET mediators

To study the individual miRNA that mediate conserved ceRNA interactions, I focused on 247 miRNAs that could be detected in all of the four datasets (Figure 4.4(A)). Surprisingly, only eleven of these were inferred by Hermes as mediating ceRNA interactions in all four networks, suggesting that conserved interactions may indeed be mediated by different miRNA in each context; see Figure 4.4(B)-(C). Because a miRNA that is in the appropriate rate-limiting kinetic regime to mediate a ceRNA interaction in one context may be expressed in a non-rate-limiting regime (i.e. too high or too low) in a different context, thus failing to provide a significant contribution.

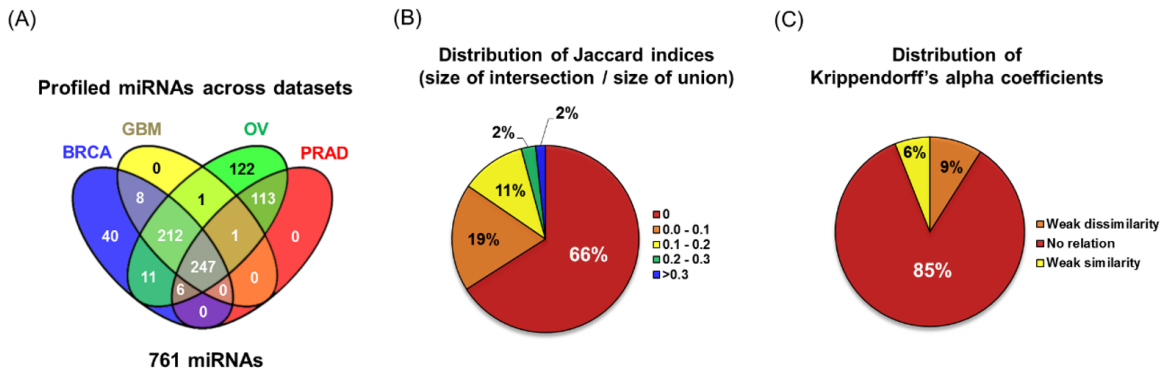


Figure 4.4. MiRNA mediators are distinct across tumor types

(A) Expression of most profiled miRNAs could be detected in at least three of the four tumor types, while expression of 247 miRNAs was detected in all four. I focus our comparative analysis on these 247 miRNAs. (B) Most miRNA mediators of PC-ceRNET interactions are context specific. For each PC-ceRNET interaction, I compared the number of mediators conserved across all four contexts to the total number of mediators with profiles in all four contexts (Jaccard index). For 66% of the interactions no common miRNA mediator was predicted in all four contexts. In total, fewer than 2% of the interactions had a Jaccard index of 30% and above. (C) Krippendorff's alpha coefficients provide an alternative measure for testing the agreement between mediation of conserved interactions across networks. Krippendorff $\alpha > 0.8$ is often used to indicate good agreement. In total, 85% of conserved interactions had $-0.15 < \alpha < 0.15$ (no relation), 9% had $\alpha < -0.15$ corresponding to very weak negative selection (weak dissimilarity), and 6% had $\alpha > 0.15$ (weak similarity). No PC-ceRNET interactions had $\alpha > 0.8$, suggesting that ceRNA mediation is not conserved across contexts.

I used the Jaccard index and Krippendorff's alpha coefficients [11] as two measures for comparing miRNA mediator sets associated with interactions in the PC-ceRNET. Compiling miRNA mediator sets for each interaction in each ceRNET, a Jaccard index was assigned by taking the ratio of the size of intersection of these sets and the size of their union. Similarly, I computed Krippendorff's alpha (α) coefficient to measure

the degree of agreement between supporting mediator sets across tumor types for each interaction in the PC-ceRNET.

Systematic analysis of PC-ceRNET interactions confirmed that the specific miRNAs inferred by Hermes to functionally mediate these interactions are highly context specific. Indeed, for 66% of the UC-ceRNA interactions, no miRNA was inferred as mediating the interaction in all four ceRNETs; see Figure 4.4(B).

To further quantify this finding using an established content-analysis standard and to evaluate its statistical significance, I computed the Krippendorff's alpha (α) coefficient for each ceRNA interaction in the PC-ceRNET. For each interaction, these coefficients describe the relative size of the pairwise overlap between its Hermes-predicted miRNAs in each of the four networks. In total, fewer than 4% of the PC-ceRNET interactions were inferred as consistently mediated by the same miRNAs across contexts, at $p < 0.05$. In addition, no interaction passed the often-used Krippendorff similarity (reliability) criteria ($\alpha > 0.8$) [12]. Indeed, the mean of the α coefficients distribution was just below 0 and 94% of the interactions had $\alpha < 0.15$ (see Figure 4.4(C)), suggesting no statistically significant overlap in the miRNA repertoire that mediates the same ceRNA interaction in different contexts. Taken together, the results suggest that an unexpectedly large proportion of ceRNA interactions is ubiquitous across cellular contexts, and yet these are mediated by different miRNAs in each context. As an example, Figure 4.5 shows predicted mediators of ceRNA interactions between *CCND1*, *HIF1A* and *HMGA2*.

Krippendorff's α coefficient is commonly used to determine the overlap between competing methods that share the same goal. The coefficient is used to decide about the reproducibility, or reliability, of these methods, and $\alpha>0.8$ is taken to indicate strong agreement, $\alpha>0.67$ is used to draw tentative conclusions, and negative values indicate disagreement. Here I compare alternate implementation of interactions in the PC-ceRNET by miRNAs. Following the procedure outlined by Krippendorff, the significance of α was obtained through an alpha distribution of a million bootstrapped samples. Using this formulation, only 3.93% of PC-ceRNET interactions were found to be consistently supported by common miRNA mediators at $p < 0.05$. The following is the description of calculating Krippendorff's alpha coefficients.

Krippendorff's α is calculated as $[1-(D_O/D_E)]$, where D_O and D_E are the observed and expected disagreements, respectively. For each interaction in the PC-ceRNET, I assemble a binary matrix A, with

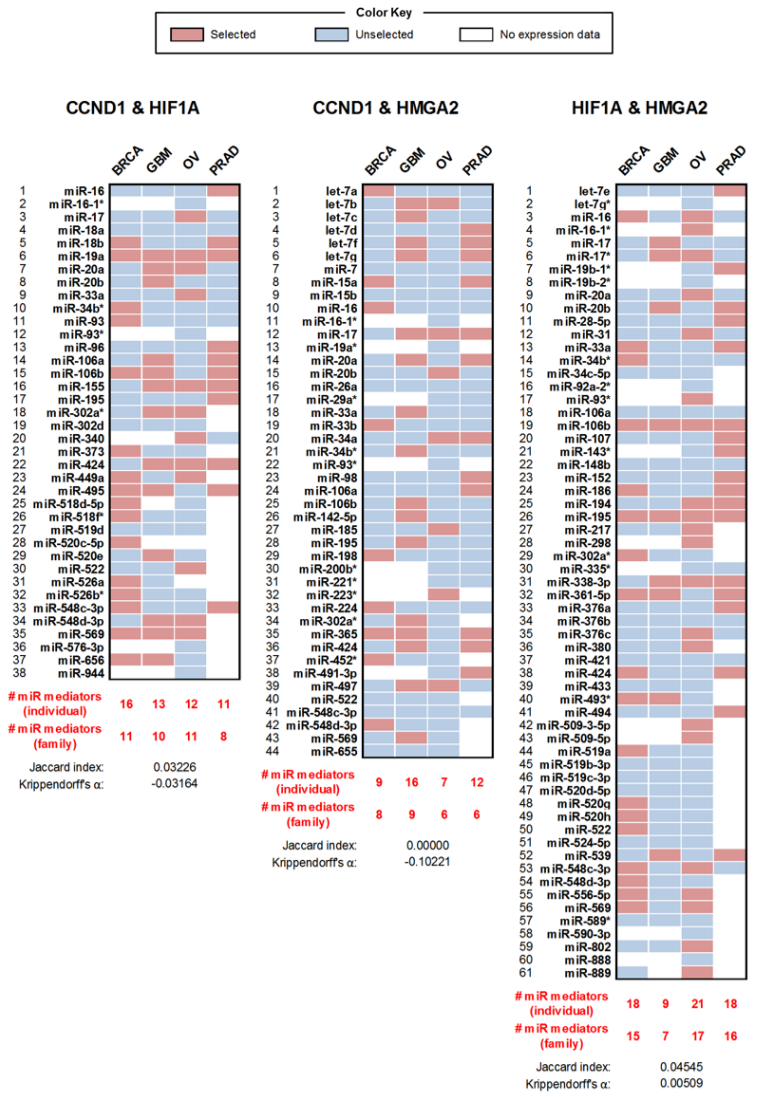


Figure 4.5. A densely connected sub-network of oncogenes in the PC-ceRNET is mediated by varying populations of miRNAs across the four tumor sets

Considering the oncogene network in Figure 4.3, I identify the miRNAs that are predicted to mediate each of the three interactions in each of the four ceRNETs. For example, 38 miRNAs were predicted to target both CCND1 and HIF1a, but their ceRNA interaction was predicted to be mediated by only 16, 13, 12 and 11 miRNAs in BRCA, GBM, OV and PRAD, respectively. miR-19a was the only mediator predicted for this interaction across all four ceRNET, resulting in a weak cross-ceRNET conservation estimate by both Jaccard index (0.03) and Krippendorff's alpha (-0.03).

missing information, to identify miRNAs that were selected as mediators of this interaction in each context. An example is given below.

	miR 1	miR 2	miR 3	miR 4	miR 5
BRCA	0	.	1	1	1
GBM	0	1	0	1	1
OV	0	1	1	1	0
PRAD	1	1	0	1	.

Matrix *A* takes on values 0 or 1, and missing information, where no profiles for the given miRNA are available, was denoted with a '·'. Using this binary matrix *A*, I construct a 2x2 coincidence matrix *O*, where each cell O_{ck} is derived from *A* by summing across miRNAs the ratio between the total number of c-k pairs and the number of tumors with available profiles for this miRNA minus 1. Here, c and k take on values 0 or 1, for with being selected or being excluded. The total number of c-k pairs observed for the miRNA is the number of ordered tumor pairs with that criteria. For example, considering miR 1 from the matrix A above, O_{01} includes the pairs (BRCA, PRAD), (GBM, PRAD) and (OV, PRAD). Similarly, O_{00} includes 6 pairs because, for example, (BRCA, GBM) and (GBM, BRCA) are both counted. Thus, summing O_{01} across columns proceeds as follows: $O_{01}=3/3+0/2+4/3+0/3+2/2$. Finally, I set $n_0= O_{00}+O_{10}$ and $n_1= O_{01}+O_{11}$, and $n=n_0+ n_1$. Then the computation of Krippendorff's*α* is given by the following formula

$$\alpha = 1 - \frac{D_o}{D_e} = \frac{(n-1)\sum_c O_{cc} - \sum_c n_c(n_c-1)}{n(n-1) - \sum_c n_c(n_c-1)}$$

To estimate significance for a given*α*, I followed Krippendorff's method for bootstrapping a distribution of *α*. Briefly, the procedure uses the original coincidence matrix to create a set of alphas with similar reliability as the derived alpha. This bootstrapped sample can then be used to test the null hypothesis that $\alpha \leq 0$.

The bootstrapping process proceeds as follows. First, define the probability of agreement $P_{agree}= (O_{00}+O_{11})/n$, then define $M = \left\{25 \sum I_{ck}, (m-1)\frac{n}{2}\right\}$, where I_{ck} is an indicator variable that is equal to 1 if $O_{ck}>$ 0, and 0 otherwise, and *m* is the total number of miRNAs in the binary matrix. Then, to obtain one bootstrapped *α*, pick *M* values in [0,1] uniformly at random and count how many of these, *q*, are larger than

P_{agree}. Finally, the bootstrapped $\alpha$is given as $\alpha=1-(q/(M\times D_e)$. After generating one million bootstrapping alphas, I count the frequency that $\alpha\leq 0$to obtain a *p*-value for rejecting the null hypothesis that $\alpha\leq 0$.

4.5 PC-ceRNET interactions are predictive of ceRNA gene expression

A unique property of ceRNETs is an increase in correlation between the expression of a specific ceRNA target and the total expression of its ceRNA regulators, as a function of the number of regulators, which has been attributed to combinatorial regulation by ceRNAs [1]. To evaluate the predictive power of PC-ceRNET interactions, I reported (1) an evaluation of the predictive ability of the PC-ceRNET on the expression of ceRNA targets in both tumor-related and non-tumor context, and (2) median correlations between ceRNA-target expression profiles and the standardized totals of the expression profiles of their predicted regulators (Figure 4.6).

I used a ridge-regression with Glmnet for Matlab within a 10-fold cross validation analysis scheme [13,14] to predict the expression of each PC-ceRNET target from the expression of its inferred ceRNA regulators. For each ceRNA target, in each 10-fold cross validation step, Glmnet constructs a regression model using training samples to fit an estimate $\hat{y}$ for ceRNA-target expression testing-sample profile y. The test-set residuals ($\hat{\varepsilon}$) are then compiled across the 10 testing-sample sets by taking the difference between the ceRNA-target expression profile y and the fitted estimate $\hat{y}$, so that $\hat{\varepsilon} = y - \hat{y}$. To calculate R^2, I take the sum of the square of the residuals across all samples, $R^2 = 1 - \sum_i \hat{\varepsilon}_i^2 \,/\, \sum_i (y_i - \bar{y})^2$, where $\bar{y}$ is the mean expression of the ceRNA target across the dataset. To assign *p*-values, to the predictive ability I used bootstrapping. Namely, the ceRNA-target expression profile y is adjusted so that $y' = \hat{y} + \delta$, where $|\delta| = |\hat{\varepsilon}|$ and δ is populated by random selection from $\hat{\varepsilon}$, with replacement. The Glmnet regression was repeated for one thousand bootstrapping y's, estimating bootstrapping R^2 using 10-fold cross validation analysis to produce a null distribution and a *p*-value was assigned by comparing the R^2 to this distribution.

Results from this analysis suggest that PC-ceRNET interactions are significantly predictive of ceRNA expression (a) in the tumor contexts from which they were inferred, (b) in other tumor contexts, and (c) even in non-tumor related contexts; see Figure 4.6. Importantly, I also observed significant correlations between a ceRNA's expression and the standardized average expression of its interacting ceRNAs; see Figure 4.6.

The results suggest that PC-ceRNET interactions contribute to gene expression regulation in a variety of cancer cells and in other pathophysiological contexts.

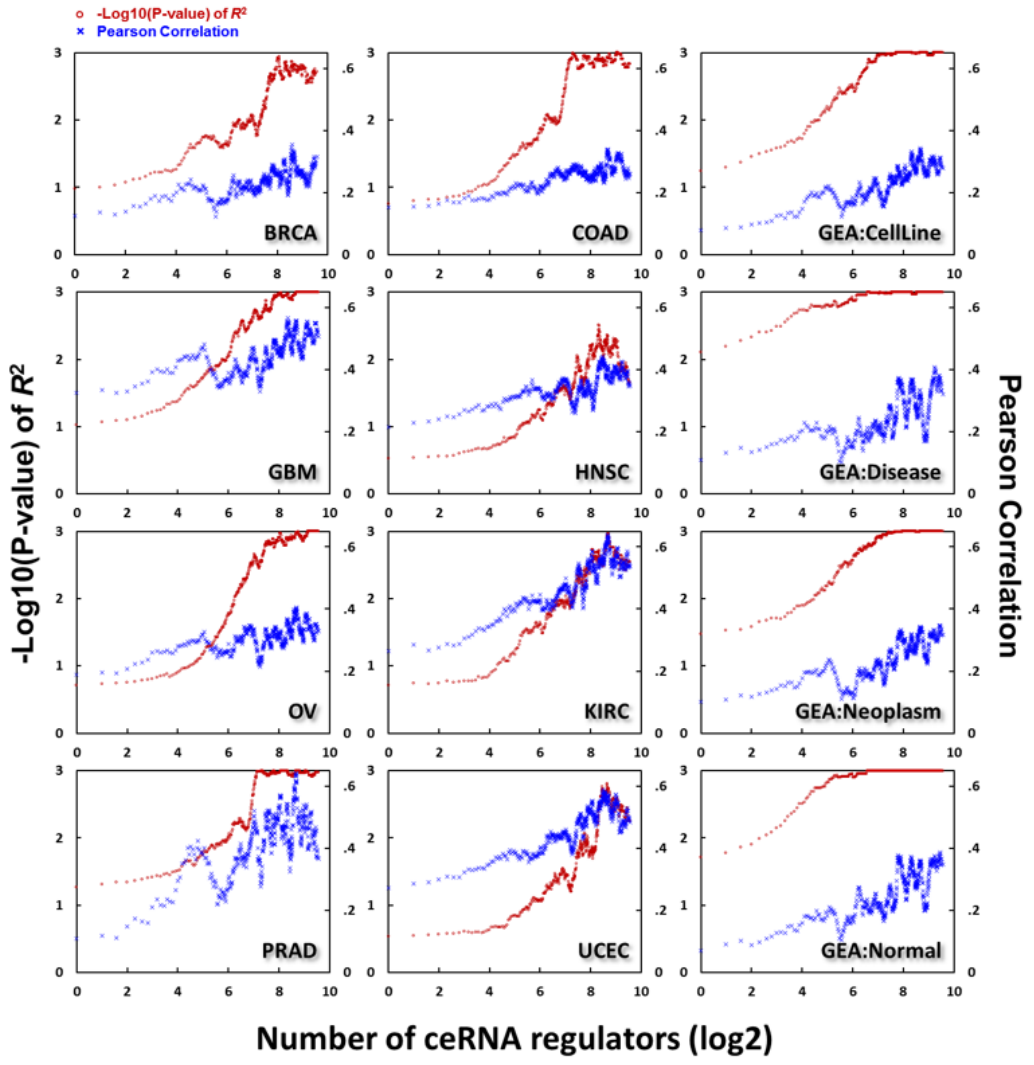


Figure 4.6. The predictive ability of the PC-ceRNET

A unique property of ceRNETs is the increase in predictive ability and correlation between expression profiles of ceRNA targets and their regulators as a function of the number of regulators (hub size). I estimated predictive ability using ridge regression analysis, assigning significance by bootstrapping residuals. In each graph, the hub size is reported on the x-axis in $\log_{10}$ scale, while the ridge-regression based p-value of the ceRNA regulators in the hub ability to predict the expression of the hub ceRNA target is reported on the y-axis. Hubs with the same number of predicted ceRNA regulators were binned together, and median *p*-values are reported for each group. As shown for virtually every context, smaller hubs are less predictive than larger hubs. In addition, I show that Pearson correlations (showing median correlation for each group) between the total standardized expression across all predicted ceRNA regulators and their targets increases as a function of the number of regulators. The left column includes datasets used to build the PC-ceRNET, the middle column includes other cancer datasets, and the right column includes cancer and non-cancer datasets provided by Gene Expression Atlas.

4.6 Summary

I predicted and worked with my colleagues to validate a pan-cancer network that includes >160,000 ceRNA interactions with evidence from both tumor and non-tumor cellular contexts. Strikingly, the majority of PC-ceRNET interactions are mediated by context-specific and disjoint miRNAs. In addition, I have proven that PC-ceRNET interactions are predictive of target gene expression even in the tumor contexts in which it was not inferred from. I will show that, in the next chapter, this network can integrate genetic and epigenetic alterations of cognate ceRNA regulators to dysregulate established oncogenes and tumor suppressors, accounting for a large fraction of the missing genomic variability in tumors.

References

1. Sumazin, P., et al., An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. Cell, 2011. 147(2): p. 370-81.

2. Darbellay , G. and I. Vajda, Estimation of the Information by an Adaptive Partitioning of the Observation Space. IEEE Trans. on Information Theory, 1999. 45: p. 1315--1321.

3. TCGA-Consortium, Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature, 2008. 455(7216): p. 1061-8.

4. Cancer Genome Atlas Research, N., Integrated genomic analyses of ovarian carcinoma. Nature, 2011. 474(7353): p. 609-15.

5. Taylor, B.S., et al., Integrative genomic profiling of human prostate cancer. Cancer Cell, 2010. 18(1): p. 11-22.

6. Buffa, F.M., et al., microRNA-associated progression pathways and potential therapeutic targets identified by integrated mRNA and microRNA expression profiling in breast cancer. Cancer Res, 2011. 71(17): p. 5635-45.

7. Lu, J., et al., MicroRNA expression profiles classify human cancers. Nature, 2005. 435(7043): p. 834-8.

8. Basso, K., et al., Reverse engineering of regulatory networks in human B cells. Nat Genet, 2005. 37(4): p. 382-390.

9. Margolin, A., et al., ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. BMC Bioinformatics, 2006. 7(Suppl 1): p. S7.

10. Jansen, R. and M. Gerstein, Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. Curr Opin Microbiol, 2004. 7(5): p. 535-45.

11. Krippendorff, K., Estimating the reliability, systematic error, and random error of interval data. Educational and Psychological Measurement, 1970. 30 (1): p. 61-70.

12. Krippendorff, K., Content analysis : an introduction to its methodology. 2nd ed. 2004, Thousand Oaks, Calif.: Sage. xxiii, 413 p.

13. Zou, H. and T. Hastie, Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005. 67(2): p. 301-320.

14. Tibshirani, R., Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc B, 1996. 58(1): p. 267-288.

Chapter 5: Identifying PC-ceRNA interactions account for missing genomic variability in tumors

5.1 Introduction

Deciphering the role of regulatory networks in propagating and coalescing the effect of multiple genomic alterations constitutes a key step towards the mechanistic understanding of complex diseases, including cancer [1,2]. Recently, multiple groups have reported that competitive endogenous RNAs (ceRNAs), both coding and non-coding, regulate each other by competing for common miRNA regulators via a simple stoichiometric mechanism [3-9]. On a genome-wide basis, I have shown that these interactions establish a regulatory layer with the potential for mediating key processes in normal cell physiology and for dysregulating them in disease. Despite these advances, however, quantitative understanding of ceRNA regulation and their overall functional relevance remain incomplete [10].

In this chapter, I will provide both computational and experimental evidence supporting the dysregulation of most oncogenes and tumor suppressors (i.e., *cancer genes*) by genetic and epigenetic alterations of their ceRNA regulators. I will focus on determining the role of the pan-cancer ceRNA network (PC-ceRNET) in tumor initiation and progression. I reasoned that ceRNA interactions may mechanistically account for a substantial fraction of the *missing genomic variability* in cancer, i.e., tumors where dysregulation of a cancer gene is not accounted for by genetic or epigenetic alterations of its locus. I thus studied whether expression of established cancer genes could be dysregulated by multiple genetic and epigenetic alterations of their predicted ceRNA regulators. My results confirm that genetic and epigenetic alterations of ceRNA regulators dysregulate hundreds of genes, including most established cancer genes, in each of eight tumor subtypes considered in this analysis. Experimentally, my colleagues show ceRNA-mediated dysregulation of *ESR1* and *APC*, an oncogene and drug target in breast cancer [11,12] and a tumor suppressor in colon adenocarcinoma [13], respectively, in samples where their loci are genetically and epigenetically intact. My colleagues then confirmed these findings by siRNA-mediated silencing of their predicted ceRNA regulators, in MCF7 breast cancer cells and HT-29 colon cancer cells, respectively.

5.2 Identification of missing genomic variability for cancer genes

Many cancer genes harbor genetic and epigenetic (genomic) alterations, which mechanistically explain their aberrant regulation in the disease [14]. Yet, cancer genes' expression is frequently dysregulated without any evidence for genomic alterations of their loci [15], a phenomenon referred to as *missing genomic variability*. In samples where genomic variability for a cancer gene is missing, its expression may be dysregulated by genomic alterations in its upstream regulators, either direct or indirect. Unfortunately, elucidating such causal distal genomic events is virtually impossible without an accurate and comprehensive map of causal regulatory interactions [2]. Based on the previous chapter, however, the PC-ceRNET provides a causal framework for identifying genomic alterations in ceRNAs that may cooperatively dysregulate the expression of specific genes of interest.

Focusing on eight tumor datasets, including glioblastoma, as well as carcinomas of the colon, head and neck, kidney, ovary, uterus, prostate, and breast, my analysis uncovered a large repertoire of ceRNAs whose genomic alterations may contribute to dysregulation of thousands of genes, including a large number of established cancer genes. First, I used elastic-net regression [16,17] with 10-fold cross validation to identify *ceRNA drivers*, whose expression is significantly predictive of the aberrant expression of genes with missing genomic variability. In total, ceRNA drivers were identified for >85% of PC-ceRNET genes missing genomic variability, across eight tumor contexts, see Figure 5.1(A).

Then, limiting my analysis to inferred driver ceRNAs and to samples where their target gene (G_{mis}) had missing genomic variability, I tested whether changes in driver ceRNAs' copy number or methylation state were statistically predictive of G_{mis} aberrant expression ($p < 0.05$). To accomplish this, I measured copy number data, CNV, as $\log_2$ ratios from aCGH and SNP array data [6], and methylation state, M, when available, as estimated by CHARM [18], from samples with missing genomic variability. I then evaluated Pearson correlations between the CNV and M profiles of drivers and G_{mis} expression profiles. Please see the following description for details.

To assess missing genomic variability for a given gene G_{mis} in the PC-ceRNET, I first selected all samples in which G_{mis}'s locus presented neither aberrant gene copy number nor differential promoter methylation. Gene are *genomically intact* in a specific sample if they have between 1.74 and 2.30 copies, as measured by sample DNA profiling, and methylation state within 2.5 standard deviations from normal tissue. To allow

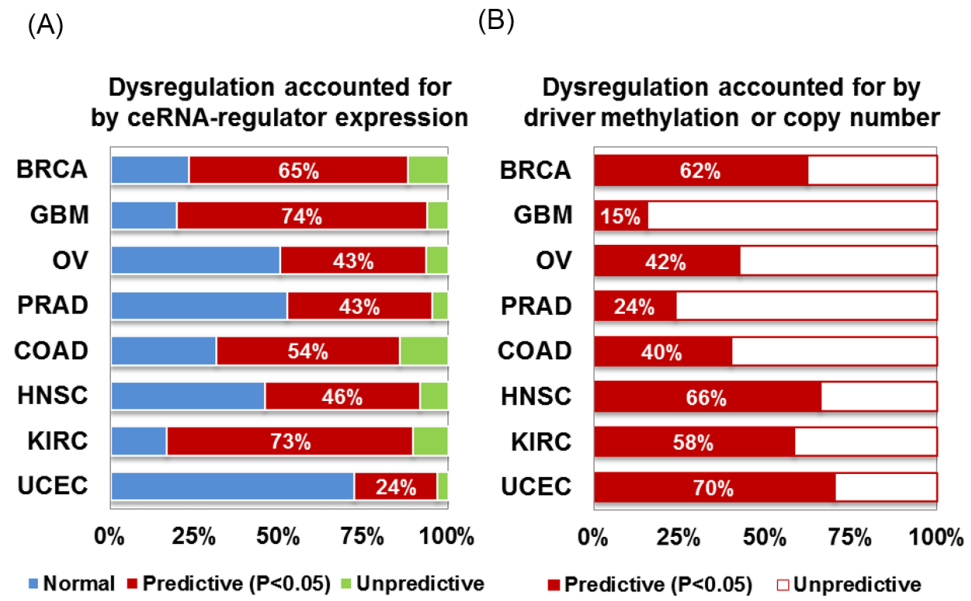


Figure 5.1. PC-ceRNET interactions account for a significant proportion of missing genomic variability across cancers.

(A) Using RNA, DNA, and methylation profiles, when available, in each of eight cancer-type datasets, I identified PC-ceRNET genes that have no genomic alterations in at least 10 samples (genomically intact samples, or gi-samples) where their expression is downregulated or upregulated compared to normal control. Blue Bar segments show the fraction of PC-ceRNET genes whose expression is not dysregulated in gi-samples. Red Bar segments show the fraction of genes whose expression is dysregulated in their gi-samples and whose expression in these samples is predicted by their ceRNA regulators ($p < 0.05$). The Green Bar segments show the fraction of genes, whose expression is dysregulated in their gi-samples and whose expression in these samples is not significantly predicted by their ceRNA regulators. (B) Focusing on ceRNA targets, whose dysregulated expression in their gi-samples can be predicted based on the expression of their ceRNA regulators, red bar segments show the fraction of genes whose expression in gi-samples is predicted by the genomic status (copy number and/or methylation) of their ceRNA regulators. White bar segments represent the fraction of genes whose expression in gi-samples is not predicted by the genomic status of their ceRNA regulators.

for appropriate cross validation testing in our analysis, only PC-ceRNET genes with at least 10 genomically intact samples were considered. Differential expression of G_{mis} in its genomically intact samples, compared to normal control samples (e.g. breast epithelium from normal biopsies in TCGA) was tested for statistical significance ($p < 0.05$, based on a two-sample Kolmogorov-Smirnov test). If the differential expression was significant, I considered the genomically intact samples to have missing genomic variability for G_{mis}, meaning that differential expression of G_{mis} in these samples could not be accounted for by genomic events at its locus. Based on this analysis, more than half of the genes represented in the PC-ceRNET genes (53%), including many cancer genes, had significant missing genomic variability. Note that focusing on

these samples excludes cases where genes are aberrantly expressed against the direction pointed to by their genomic alterations; for example, genes with amplified loci may be downregulated relative to normal samples. While my procedure for identifying genes with missing genomic variability may be altered to explicitly account for these cases and the multitude of their subcases, the prevalence of these cases in our data is low, making for less than 10% of total.

Note that for prostate cancer and glioblastoma, only copy number data could be used, since methylation data is not available in the prostate cancer datasets and the normal methylation baseline is not available in glioblastoma.

Significant correlation in either test was accepted as evidence that genomic alterations of ceRNA drivers are predictive of their target's expression. Across all tested tumor types, our analysis shows that genomic alterations of inferred ceRNA drivers mechanistically account for a significant fraction of the aberrant expression of almost 50% of PC-ceRNET genes with missing genomic variability, see Figure 5.1(B). Note that driver ceRNAs were inferred purely based on whether their expression was predictive of the expression of a target ceRNA with missing genomic variability.

5.3 PC-ceRNET accounts for missing genomic variability of cancer genes

To identify drivers of established cancer genes, I compiled a list of 2,040 genes previously associated with cancer initiation and progression, from the Cancer Gene Census [19], the CancerGenes resource [20], Fred Waldman's cancer gene set, the Tumor-Associated Gene Database [21], and Cen et. al. [22]. Of these, 839 are represented in the PC-ceRNET, where they are regulated by an average of 101 ceRNAs. 212 of the 839 present focal, recurrent locus alterations in at least one tumor type and are implicated in cancer initiation by multiple lines of evidence [20]. I thus focused on this 212-gene repertoire, including extensively characterized cancer genes, such as: *APC*, *BCL2*, *Cyclin D1*, *DICER*, *EGFR*, *HMGA2*, *IDH1*, *KRAS*, *MET*, *MYB*, *MYC*, *NOTCH*, *PBX1*, *PDGFRA*, *PIK3*, *PTEN*, *RB1*, *RUNX1*, *TP53*, and *TP63*.

Table 5.1 provides a summary of key cancer genes, for each of the eight tumor types, whose dysregulation in their genomically intact samples is predicted by genomic alterations of their driver ceRNAs. Of 212 cancer genes with recurrent mutations, the vast majority (n = 179) were predicted to be significantly dysregulated by genomic alteration of their predicted ceRNA drivers in at least one tumor context where they had missing genomic variability. Moreover, missing genomic variability for 44 of these 179 genes, including *CCND2*,

DICER, *IDH1*, *KIT*, *KRAS*, *MYCN*, *NCOA1*, *NFIB*, *NRAS*, *PDGFRA*, *PTEN*, and *RUNX1* could be accounted for by genomic alterations of their predicted ceRNA drivers in at least four of the tumor contexts considered in this analysis. This suggests that modulation of these genes by their cognate ceRNA regulators is a common event. In total, of the 839 cancer genes in the PC-ceRNET, 682 presented alterations of their drivers that were predictive of their expression in at least one tumor context, and 512 in two or more tumor contexts.

BRCA	GBM	OV	PRAD	COAD	HNSC	KIRC	UCEC
BCL2 (5%)	CCND1 (86%)	CCND2 (3%)	AR (84%)	ACVR2A (59%)	E2F1 (18%)	BCL2 (5%)	DICER1 (23%)
CBFB (5%)	CEBPB (70%)	CDKN1B (16%)	CCND2 (90%)	APC (19%)	EGFR (12%)	EGR1 (22%)	E2F1 (10%)
CCNE2 (33%)	EGFR (14%)	DDX5 (15%)	FGFR2 (90%)	AR (12%)	MYC (21%)	HIF1A (29%)	NRAS (48%)
CDC42 (38%)	IDH1 (96%)	E2F1 (20%)	FOS (88%)	CCND2 (18%)	NRAS (45%)	KRAS (35%)	PTEN (8%)
CDKN1B (52%)	IGF2BP3 (23%)	IGF2BP2 (3%)	PDGFRA (91%)	CDKN1A (52%)	RB1 (6%)	MAML1 (14%)	SOX4 (35%)
EGR1 (42%)	MET (21%)	KIT (17%)	PIK3R3 (94%)	CDKN1B (54%)	RHOA (20%)	MET (26%)	
ESR1 (9%)	PDGFRA (78%)	KRAS (6%)	QKI (93%)	DICER1 (29%)	SMAD4 (27%)	PBRM1 (7%)	
FOS (46%)	RUNX1 (89%)	MAPK1 (6%)	STAT3 (94%)	EPHA4 (52%)	TP53 (28%)	PDGFRA (23%)	
GATA3 (39%)	SMAD2 (84%)	NRAS (24%)	TP63 (93%)	EPHB2 (56%)	TP63 (12%)	PTEN (15%)	
HMGA2 (12%)	SMAD4 (85%)	PTEN (9%)	VCL (93%)	KRAS (43%)	VEGFA (39%)	SETD2 (6%)	
MAP3K1 (45%)		RBL2 (13%)		MYC (35%)		VEGFA (32%)	
NF1 (23%)		TP53 (12%)		NRAS (50%)			
NRAS (54%)		WEE1 (13%)		SMAD4 (27%)			
PIK3CA (44%)				ZEB2 (24%)			
PIK3R1 (27%)							
RUNX1 (18%)							

Table 5.1. Missing genomic variability of key cancer genes recovered by alterations at their ceRNA regulators

For each tumor dataset, including breast cancer, glioblastoma, and carcinomas of the ovary, prostate, colon, head and neck, kidney and uterus, I list known oncogenes and tumor suppressors whose unexplained dysregulation is predicted by alterations at their ceRNA regulators. The proportion of samples with intact loci but aberrant expression is given in parentheses.

As illustrative examples, I provide detailed experimental analysis of *ESR1*, a breast cancer oncogene, and of *APC*, a tumor suppressor in colon adenocarcinoma. Considering TCGA breast cancer data, *ESR1* was found over-expressed in 46 breast-cancer samples with a genomically intact locus, compared to normal breast epithelial tissue, see Figure 5.2(B). Conversely, *APC* was under-expressed in 27 colon-cancer

samples with a genomically intact locus, compared to normal colon tissue; see Figure 5.3(B). I identified ten ceRNA drivers for ESR1 and nine for APC for which (a) total gene expression, (b) total copy number, and (c) total methylation status were predictive of *ESR1* and *APC* expression in samples with missing genomic variability, respectively; see Figure 5.2(A) and 5.3(A).

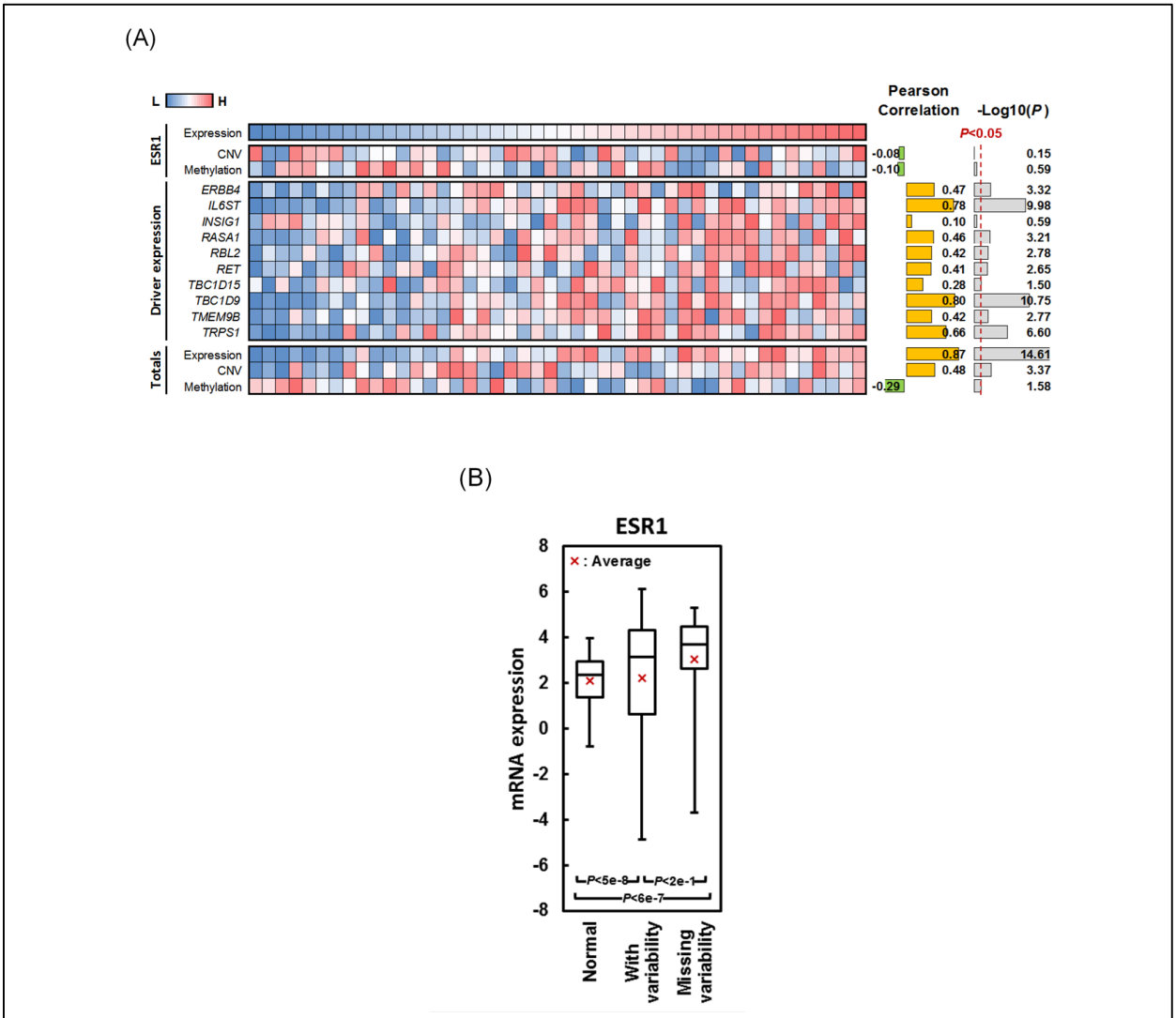


Figure 5.2. ESR1-regulating ceRNA drivers.

(A) DNA and methylation profiles of the *ESR1* loci are not predictive of their expression. However, expression of individual ceRNA drivers of *ESR1* are mostly correlated with their expression, and standardized total expression profiles, DNA copy number profiles, and methylation profiles of ceRNA drivers *of ESR1* are significantly correlated with their expression. (B) *ESR1* is significantly over-expressed in breast cancer gi-samples.

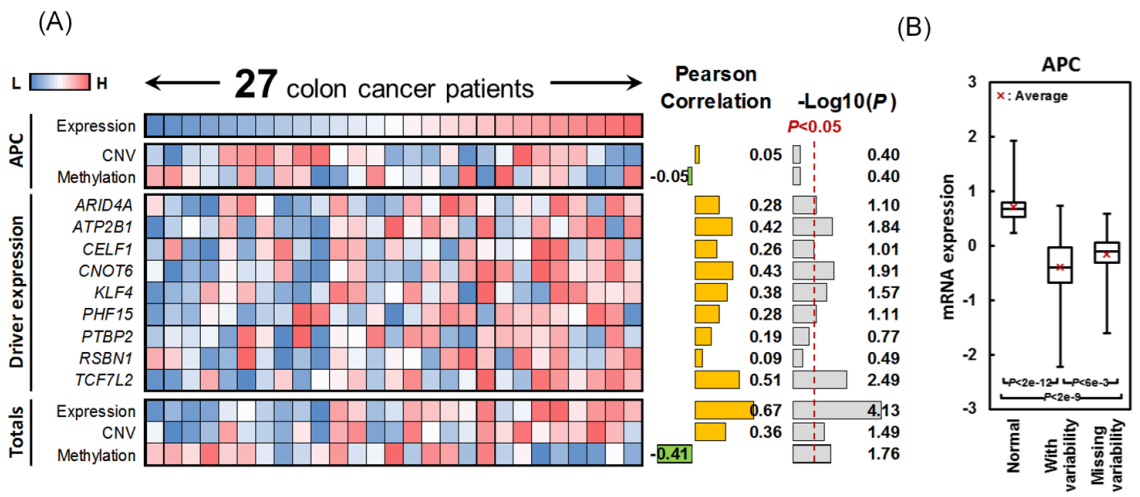


Figure 5.3. APC-regulating ceRNA drivers

(A) DNA and methylation profiles of the *APC* loci are not predictive of their expression. However, expression of individual ceRNA drivers of *APC* are mostly correlated with their expression, and standardized total expression profiles, DNA copy number profiles, and methylation profiles of ceRNA drivers *of APC* are significantly correlated with their expression. (B) *APC* is significantly under-expressed in colon cancer gi-samples.

siRNA-mediated silencing of the vast majority of these ceRNA drivers in MCF7 luminal breast cancer cells and in HT-29 colon adenocarcinoma cells consistently and significantly reduced 3'-UTR luciferase activity of *ESR1* and *APC*, respectively (Figure 5.4). Co-silencing of ceRNA driver pairs that are co-amplified or co-deleted in patients and whose total expression improves correlation with *ESR1* and *APC* expression in tumor samples further decreased *ESR1* and *APC* 3'-UTR luciferase activity (Figure. 5.4(B) and 5.4(D)). On average, siRNA mediated silencing of single ceRNA drivers and of driver-pairs decreased ESR1 3'-UTR luciferase activity by 14% and 22%, respectively. For *APC* the proportions were 9% and 21%, respectively. Thus, these experiments support the additive cooperative regulation activity by ceRNA driver pairs. In the following, I will describe all related steps in identifying ceRNA drivers.

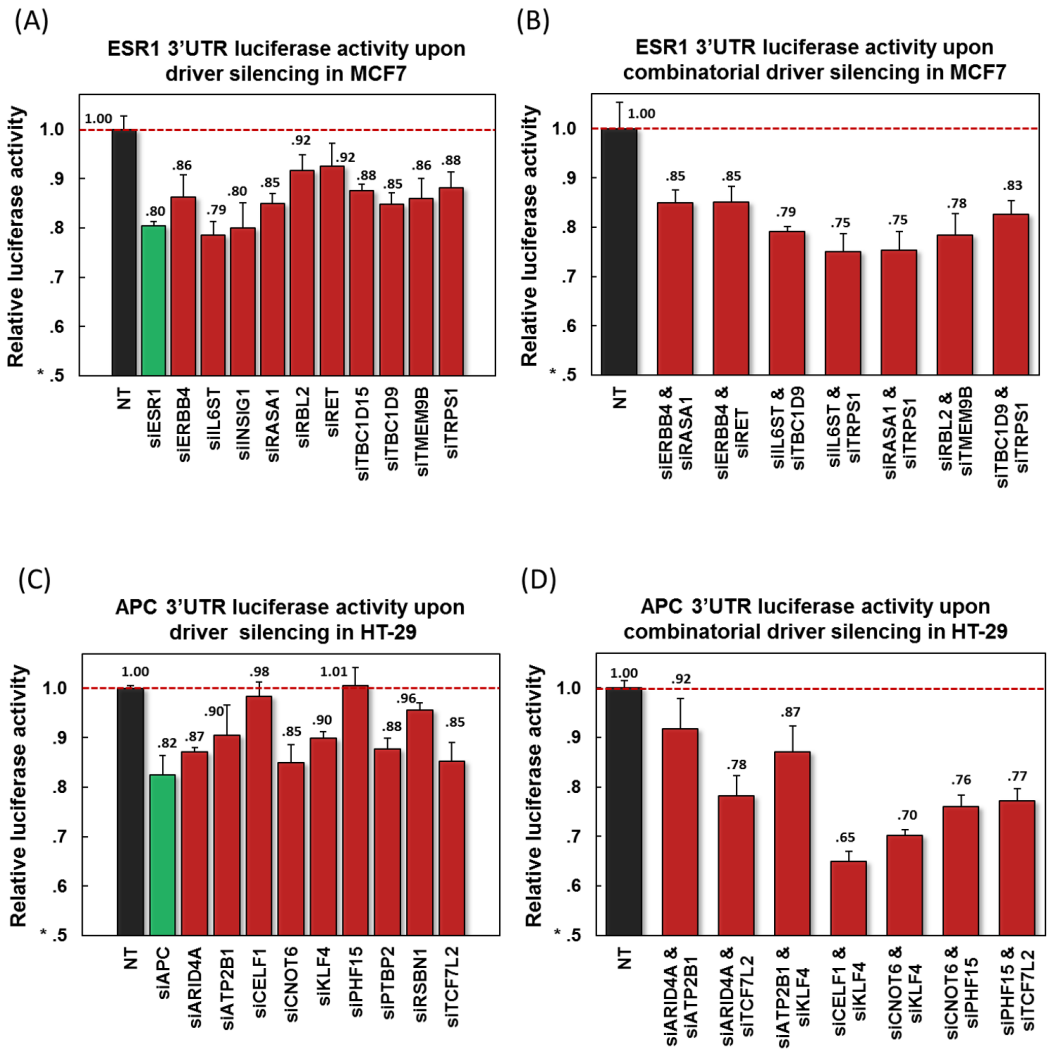


Figure 5.4. Biochemical validation of regulation by ceRNA drivers

(A) *ESR1* 3' UTR luciferase activity is downregulated by single silencing of drivers and (B) combined silencing of driver-pairs in MCF7. (C) *APC* 3' UTR luciferase activity is downregulated by single silencing of drivers and (D) combined silencing of driver-pairs in HT-29. Combined driver pairs were selected to correspond to co-altered loci in the breast and colon cancer datasets. Data are represented as mean ± SEM.

5.4 Selecting candidate drivers

Given expression profiles for a ceRNA target and its *N* predicted ceRNA regulators, I selected candidate drivers by first clustering them according to their expression profiles. Clustering was performed using *k*-means with all possible choices for *k*, where each cluster is represented by its centroid. Then, for each *k*, elastic net regression and 10-fold cross validation was used to estimate a test-set residual sum of squares

and the corresponding Akaike information criterion (AIC) [23]. Note that elastic net regression is commonly used for identifying interactions [17], AIC is a distance measure that punishes likelihood functions that are based on more variables, and 10-fold cross validation was used here to rank and select solutions rather than evaluate their overall significance. Genes contributing to at least 50% of the top $\sqrt{N}$ results by AIC, after sample size correction, were selected as candidate drivers. Finally, I summed across standardized expression profiles of candidate drivers and compared the correlation between these total profiles and the expression profile of the target gene. To assign significance for this final selection of ceRNAs, I compared the resulting correlation coefficient to a distribution of correlations obtained by shuffling sample labels (selecting $p < 0.05$).

I clustered ceRNA regulators and represented them by cluster centroids (super genes) to improve prediction rates and aid in significance testing [24], while allowing for the inclusion of correlated ceRNA regulators during candidate driver selection following regression. Specifically, elastic net regression produces regression models with sparse variable selections. By representing correlated genes as aggregate variables I reduce the number of variables for selection by elastic net regression while ensuring that correlated drivers, which may be omitted by elastic net regression because their simultaneous inclusion in a predictive model does not improve the fit, could be considered when making candidate driver selections. Thus, after centroid selection by elastic net regression all represented ceRNA regulators are considered in the next selection step.

Similarly to the procedure described for evaluating the predictive power of PC-ceRNET interactions, regression with10-fold cross validation was used to estimate the reduction in variance. Then Akaike information criterion (AIC) was computed as $\mathrm{AIC} = n \ln(\sum_i \hat{\varepsilon}_i^2 / n) + 2k$, where n is the number of samples in the dataset and k is the number of clusters used. To correct for small sample size and avoid overfitting the data, especially in cases where n is small relative to k, I used the sample-size correction $\mathrm{AIC}' = \mathrm{AIC} + \frac{2k(k+1)}{n-k-1}$ [25]. This criterion was used to compare regression models across k, where lower AIC' is associated with improved sample-size corrected predictive power. Expanding centroid to genes they represent, selecting as candidate drivers those ceRNA regulators that that contributed to at least 50% of the top $\sqrt{N}$ results by AIC'.

5.5 Selecting driver pairs for validation

I selected ceRNA pairs based on two criteria. First, the loci of both ceRNA must be amplified (ESR1) or deleted (APC) at aCGH log ratio >0.2 or < -0.2, respectively, in at least one of the samples with missing genomic information. Second, the sum of standardized expression profiles must have higher correlation with the ESR1/APC expression profiles than the individual profile of the two contributing drivers across the samples.

5.6 Selecting drivers and evaluating predictive power using regression

I used elastic net regression to derive and evaluate predictive models, to estimate the predictive power of ceRNA regulator expression, and to identify candidate drivers. Elastic net regression combines L1 and L2 regularized regression penalty terms in order to strike a balance between obtaining a parsimonious model (through the L1 term), while retaining groups of correlated features (through the L2 term). As input, the algorithm took on standardized expression profile matrices $X$for ceRNA regulators which were used to predict the standardized expression profile of their target$y$. I used Glmnet for Matlabto solve the following optimization problem:

$$l(\beta)_{penalized} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{sample\ i} \left(y_i - \beta_0 - \sum_{regulator\ j} \beta_j X_{ij} \right)^2 + \lambda_1 ||\beta|| + \lambda_2 ||\beta||^2 \right\}$$

$l(\beta)$is the regression model with coefficients β, and $\left|\left|\beta\right|\right|$ denotes the inner product of β; β_0 is called the constant or the intercept. The penalty for selected model coefficients is determined by $\lambda_1$and λ_2. When searching to minimize $l(\beta)_{penalized}$, I fix $\alpha = \lambda_2/(\lambda_2 + \lambda_1)$ and use Glmnettosimultaneously search for a legal $(\beta, \lambda_1, \lambda_2)$ combination. The process is repeated for $\alpha \in [0.1, 1]$followinga0.1 increment. The best α was chosen to minimize $l(\beta)_{penalized}$ according to 10-fold cross validation. When computing ridge regression, β has no zero entries and I set $\lambda_1 = 0$ (eliminating L1 penalty).

5.7 Summary

Following elucidation of ceRNA interactions [3-5] and discovery of a large-scale ceRNA regulation layer in glioblastoma [6] intense debate has ensued about the role that ceRNAs may play in controlling normal cell physiology and disease [10]. For instance, it has been suggested that since highly expressed miRNAs are

much more abundant than their cognate mRNA targets, they almost never operate in a rate-limiting regime and thus pathophysiologically relevant ceRNA regulation may be a rare event [10]. While this is indeed the case for ceRNA interactions mediated by a single miRNA, I also show that mediation by multiple miRNAs induces significant ceRNA coupling efficiency, virtually independent of individual miRNA expression. As a result, my morel predicts that ceRNA interactions mediated by many miRNAs should be conserved across cellular contexts. Finally, the kinetic model shows that cooperative regulation of a target ceRNA by several ceRNA regulators is roughly additive (i.e., linear), suggesting that genomic dysregulation of multiple ceRNAs, all cooperatively regulating the same disease-relevant target ceRNA, may be an important mechanism of disease initiation and progression and explain some of the missing genomic variability, including in cancer.

Supporting the model predictions, Hermes-predicted and experimentally validated ceRNA interactions share an average of 17 miRNAs, thus suggesting that a majority of these should be pathophysiologically relevant and should be furthermore conserved across different cellular contexts. An important corollary of these observations is that mutation of a single miRNA binding site should likely have no appreciable effect on the mediated ceRNA interaction because the effect is mediated by a large number of distinct miRNA species. This is both important in the context of ceRNA interaction validation and intriguing from an evolutionary standpoint as it suggests that ceRNA interactions are built to be highly insensitive to 3'UTR mutations.

Finally, I show that the PC-ceRNET provides an effective mechanism for dysregulating the expression of a significant fraction of cancer genes in samples where their loci are genomically intact. Taken together, these findings suggests that ceRNA interactions constitute an integral, highly functional, and unusually robust component of the cell's regulatory machinery and that their genomic alteration may have important functional consequences in tumorigenesis and in the etiology of other disease.

My analysis extends previous reports suggesting that genomic alteration of a few ceRNAs may be linked to disease initiation to a genome-wide scale, suggesting that an unexpectedly high fraction of the missing genomic variability in cancer may be mechanistically accounted for by alterations of ceRNA regulators of key cancer genes. Thus, ceRNA regulation constitutes an important mechanism to help elucidate the full

repertoire of genetic and epigenetic alterations contributing to tumorigenesis and tumor progression that cannot be identified by statistical analysis due to their cooperative effect [2].

Finally, I showed that conserved ceRNA interactions are predictive of gene expression even in non-tumor related contexts, suggesting that this new extensive regulatory layer may play an important role in normal cell physiology and, in particular, in maintaining cellular homeostasis. Indeed, dysregulation of the ceRNA regulatory layer, may have relevant effects in complex diseases other than cancer, where the homeostatic machinery of the cell is compromised, such as in diabetes and obesity, for instance.

Taken together, these results suggest that the PC-ceRNET is a novel and valuable resource for the study of cancer and of other disease and that its further understanding may elucidate critical mechanisms implemented by this regulatory layer to control normal cell physiology.

References

1. Cowin, P.A., et al., Profiling the cancer genome. Annu Rev Genomics Hum Genet, 2010. 11: p. 133-59.

2. Califano, A., et al., Leveraging models of cell regulation and GWAS data in integrative network-based association studies. Nat Genet, 2012. 44(8): p. 841-7.

3. Franco-Zorrilla, J.M., et al., Target mimicry provides a new mechanism for regulation of microRNA activity. Nat Genet, 2007. 39(8): p. 1033-7.

4. Lee, D.Y., et al., Expression of versican 3'-untranslated region modulates endogenous microRNA functions. PLoS ONE, 2010. 5(10): p. e13599.

5. Poliseno, L., et al., A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature, 2010. 465(7301): p. 1033-8.

6. Sumazin, P., et al., An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. Cell, 2011. 147(2): p. 370-81.

7. Cazalla, D., T. Yario, and J.A. Steitz, Down-regulation of a host microRNA by a Herpesvirus saimiri noncoding RNA. Science, 2010. 328(5985): p. 1563-6.

8. Tay, Y., et al., Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. Cell, 2011. 147(2): p. 344-57.

9. Cesana, M., et al., A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. Cell, 2011. 147(2): p. 358-69.

10. Ebert, M.S. and P.A. Sharp, Roles for MicroRNAs in Conferring Robustness to Biological Processes. Cell, 2012. 149(3): p. 515-24.

11. van 't Veer, L.J., et al., Gene expression profiling predicts clinical outcome of breast cancer. Nature, 2002. 415(6871): p. 530-6.

12. Cancer Genome Atlas, N., Comprehensive molecular portraits of human breast tumours. Nature, 2012. 490(7418): p. 61-70.

13. Cancer Genome Atlas, N., Comprehensive molecular characterization of human colon and rectal cancer. Nature, 2012. 487(7407): p. 330-7.

14. Forbes, S.A., et al., The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr Protoc Hum Genet, 2008. Chapter 10: p. Unit 10 11.

15. Akavia, U.D., et al., An integrated approach to uncover drivers of cancer. Cell, 2010. 143(6): p. 1005-17.

16. Zou, H. and T. Hastie, Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005. 67(2): p. 301-320.

17. Barretina, J., et al., The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature, 2012. 483(7391): p. 603-7.

18. Irizarry, R.A., et al., Comprehensive high-throughput arrays for relative methylation (CHARM). Genome Res, 2008. 18(5): p. 780-90.

19. Futreal, P.A., et al., A census of human cancer genes. Nat Rev Cancer, 2004. 4(3): p. 177-183.

20. Higgins, M.E., et al., CancerGenes: a gene selection resource for cancer genome projects. Nucleic Acids Res, 2007. 35(Database issue): p. D721-6.

21. Galperin, M.Y. and X.M. Fernandez-Suarez, The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. Nucleic Acids Res, 2012. 40(Database issue): p. D1-8.

22. Chen, J., et al., Genomic profiling of 766 cancer-related genes in archived esophageal normal and carcinoma tissues. Int J Cancer, 2008. 122(10): p. 2249-54.

23. Akaike, H., A new look at the statistical model identification. IEEE Transactions on Automatic Control, 1974. 19(6): p. 716–723.

24. Park, M.Y., T. Hastie, and R. Tibshirani, Averaged gene expressions for regression. Biostatistics, 2007. 8(2): p. 212-27.

25. Burnham, K.P., D.R. Anderson, and K.P. Burnham, Model selection and multimodel inference : a practical information-theoretic approach. 2nd ed. 2002, New York: Springer. xxvi, 488 p.